Ordnende Prinzipien statistischer Korrelationen in eukaryotischen Genomen

Dem Fachbereich Biologie der Technischen Universität Darmstadt
zur Erlangung des akademischen Grades eines
Doctor rerum naturalium (Dr. rer. nat.)
vorgelegte Dissertation von

Manuel Dehnert

aus Bad Hersfeld

1. Referent: Prof. Dr. Marc-Thorsten Hütt
2. Referent: Prof. Dr. Felicitas Pfeifer
3. Referent: Prof. Dr. Werner E. Helm

Tag der Einreichung: 22. Mai 2006
Tag der mündlichen Prüfung: 30. Juni 2006

Darmstadt 2006
D17

Die vorliegende Arbeit wurde am Institut für Botanik der Universität Darmstadt in der Arbeitsgruppe Bioinformatik von Herrn Prof. Dr. Marc-Thorsten Hütt in der Zeit von Februar 2003 bis Mai 2006 angefertigt.

Publikationen

Teile dieser Doktorarbeit sind in folgende Publikationen eingegangen:

Dehnert, M., Helm, W.E. und Hütt, M.-Th., 2003. A discrete autoregressive process as a model for short-range correlations in DNA sequences. Physica A 327, 535-553.

Dehnert, M., Helm, W.E. und Hütt, M.-Th., 2005. Information theory reveals large-scale synchronisation of statistical correlations in Eukaryote genomes. Gene, 345, 81-90.

Dehnert, M., Plaumann, R., Helm, W.E. und Hütt, M.-Th., 2005. Genome phylogeny based on short-range correlations in DNA sequences, J. Comp. Biol. 12, 545-553.

Dehnert, M., Helm, W.E. und Hütt, M.-Th., 2006. The informational structure of two closely related eukaryotic genomes. Phys. Rev. E, eingereicht.

Hütt, M.-Th. und Dehnert, M., 2006. Methoden der Bioinformatik. Eine Einführung. Springer-Verlag, Heidelberg, Berlin.

Beiträge zu Konferenzen

Structural approaches to sequence evolution: Molecules, networks, populations
Max-Planck-Institut für Physik komplexer Systeme, Dresden 2004
Poster: *Method for visualising the robustness of phylogenetic trees*

German Conference on Bioinformatics 2004
Universität Bielefeld, Bielefeld 2004
Vortrag: *Repetitive sequence elements explain only part of the phylogenetic information in the correlation structure of Eukaryote genomes*

69. Jahrestagung der Deutschen Physikalischen Gesellschaft
Humboldt-Universität zu Berlin und
Technische Universität Berlin, Berlin 2005
Poster: *Information theory reveals large-scale synchronisation of statistical correlations in Eukaryote genomes*

Inhaltsverzeichnis

Einleitung

Die Analyse statistischer Eigenschaften von DNA-Sequenzen ist seit mehreren Jahrzehnten ein recht großes interdisziplinäres Forschungsfeld, in dem Wissenschaftler mit ganz unterschiedlicher fachlicher Ausbildung einen Beitrag zum besseren Verständnis biologischer Sachverhalte liefern. In besonderem Maße zeichnet ein Wissenstransfer durch Anwendung von Methoden aus den Bereichen der Mathematik und statistischen Physik diese Forschungsrichtung aus. Korrelationen in DNA-Sequenzen, also statistische „Abhängigkeiten“ innerhalb der Sequenzen, haben sich als ein sehr guter Zugang erwiesen, um biologische Eigenschaften zu quantifizieren und zu beschreiben.

Solch eine ganz offensichtliche biologische Eigenschaft eines Organismus ist seine evolutionäre Entwicklung und Differenzierung. Korrelationen unmittelbar benachbarter Basen, *nearest neighbor base-base correlations*, führen auf ein charakteristisches Wertemuster für eine Spezies (Russell et al., 1976; Russell und Subak-Sharpe, 1977). In Bezug auf eine mögliche phylogenetische Interpretation sind die Arbeiten von Karlin und Ladunga (1994), Karlin und Mrázek (1997) und Gentles und Karlin (2001) von besonderer Bedeutung. Sie zeigen, dass die Verteilung von Dinukleotiden (also Nukleotidpaaren) bei prokaryotischen und eukaryotischen Organismen speziesabhängig ist und eine Genom-Signatur darstellt. In den Untersuchungen von Karlin et al. wird vermutet, dass die beobachteten Unterschiede zwischen Spezies auf der Dinukleotidebene mit DNA-Replikation und DNA-Reparaturmechanismen in Verbindung stehen. In neueren Arbeiten werden Häufigkeitsunterschiede auf nachbarschaftsabhängige Mutationsraten zurückgeführt (Arndt et al., 2002; Arndt und Hwa, 2005). Auf Basis dieses mathematischen Modellansatzes lassen sich beobachtete Unterschiede in den Häufigkeiten von Nukleotiden und Dinukleotiden durch unterschiedliche Mutationsraten bei verschiedenen Spezies erklären. Von den 1990er Jahren an bis heute sind auf der Grundlage und unter Verwendung weiterer statistischer Sequenzeigenschaften Genom-Signaturen formuliert worden. Dabei wurde der Ansatz der Zwei-Wort (Dinukleotid-)Verteilungen auch auf die Betrachtung von n-Wort Verteilungen übertragen (Hao und Qi, 2003; Qi et al., 2004).

Als sehr erfolgreich hat sich die Betrachtung von Korrelationen über einen Symbolabstand k erwiesen. Damit wird ein Übergang zu größeren Skalen ermöglicht und es lässt sich eine Verbindung von statistischen mit strukturellen Eigenschaften der DNA-Sequenzen herstellen. Ein relativ einfacher Zusammenhang zwischen Struktur und Korrelation ergibt sich für proteincodierende DNA-Sequenzen. Die Triplet-Struktur in Form von Codons führt zu Periode-3-Oszillationen in einem geeigneten Korrelationsmaß, z.B. der Transinformation (Herzel und Grosse, 1997; Grosse et al., 2000). Der Grund dafür liegt in den unterschiedlichen Wahrscheinlichkeiten für das Auftreten der Basen in jeder Position eines Codons, verbunden mit der Nicht-Gleichverteilung dieser Basen in-

nerhalb von codierenden Sequenzbereichen (*nonuniform codon usage*). Korrelationen im Abstand zwischen 10 und 11 Basen werden mit der DNA-Faltung innerhalb der Nukleosomen assoziiert (Trifonov und Sussman, 1980; Herzel et al., 1999). Die interne Struktur einer Klasse von repetitiven Elementen im menschlichen Genom, den *Alu*-Repeats, führt zu einem deutlichen Signal in der Transinformation bei Symbolabständen zwischen 100 und 200 Basen (Holste et al., 2003).

Es wird eine kontroverse Debatte darüber geführt, wie statistische Abhängigkeiten in DNA-Sequenzen, die über mehrere Größenordnungen existieren, zu erklären sind. Die als langreichweitige Korrelationen bezeichneten Abhängigkeiten wurden erstmals Anfang der 1990er Jahre in DNA-Sequenzen nachgewiesen (Li und Kaneko, 1992; Peng et al., 1992; Voss, 1992). Als mögliche Ursache dafür werden unterschiedliche biologische Eigenschaften von DNA-Sequenzen diskutiert. Große Aufmerksamkeit wurde der These zuteil, dass die mosaikhafte Struktur von DNA-Sequenzen (Bernardi et al., 1985) für die beobachteten langreichweitigen Korrelationen verantwortlich sei (Maddox, 1992; Karlin und Brendel, 1993; Chatzidimitriou-Dreismann und Larhammar, 1993; Peng et al., 1994). Die Auswirkung der Verteilung biologisch klar benennbarer Komponenten in DNA-Sequenzen bildet einen weiteren Ansatz, um die beobachteten langreichweitigen Korrelationen zu erklären. Dazu gehört die Längenverteilung proteincodierender Segmente (Herzel und Grosse, 1997) und ihre Alternation mit nicht-codierenden Sequenzabschnitten charakteristischer Länge (Nee, 1992). Es wurde untersucht, ob die Verteilung von Retrotransposons einen Beitrag zu diesen Korrelationen leistet (Holste et al., 2003) und ob Variationen im GC-Gehalt entlang der Sequenz mit langreichweitigen Korrelationen in Verbindung stehen (Li und Holste, 2004b,a, 2005). Nach heutigem Stand erklärt keiner der diskutierten Ansätze die beobachteten langreichweitigen Korrelationen schlüssig (Li und Holste, 2005), auch wenn Modelle der Sequenzevolution zum Verständnis dieses Phänomens beigetragen haben (Li, 1989, 1991; Messer et al., 2005). Hier wird sehr deutlich, wie gut solche Korrelationsanalysen Sequenzeigenschaften kondensiert zusammenfassen (z.B. die Periode-3-Oszillationen für codierende Bereiche) und welche Schwierigkeiten man zu überwinden hat, bis diese abstrakten Befunde mit konkreten Phänomenen in Verbindung gebracht werden können (z.B. im Fall langreichweitiger Korrelationen).

Die Analyse von Korrelationen in DNA-Sequenzen über einen größeren Symbolabstand bildet die Grundlage der vorliegenden Arbeit. Am Anfang steht die Formulierung einer neuen Genom-Signatur, die durch ein starkes, innerhalb der Chromosomen einer Spezies hoch synchronisiertes Signal gebildet wird. Diese Signatur basiert auf kurzreichweitigen Korrelationen von Basen in DNA-Sequenzen. Das Interesse der Untersuchung liegt im Weiteren in der Verbindung dieser statistischen Eigenschaften mit ihren biologischen Ursachen. Dabei wird der Versuch unternommen, Spezies-Information mit funktionell benennbaren Elementen der DNA in Verbindung zu setzen. Dieses Vorgehen wird ermöglicht durch die Anwendung einer neuen Schätzmethode für die Stärke der Korrelation in einem größeren Symbolabstand, die eine Subtraktion des Rauschanteils innerhalb der DNA-Sequenzen erlaubt.

Ein Schwerpunkt liegt auf dem Beitrag repetitiver DNA zu dieser neuen Genom-Signatur. Damit stellt die Arbeit einen Brückenschlag dar zwischen dem Forschungsgebiet der Analyse von Korrelationen in DNA-Sequenzen und der Betrachtung von Speziesunterschieden in Prozessen der Genom-Evolution.

1

Mathematische Methoden

Die in dieser Arbeit eingesetzten mathematischen Methoden haben ihren Ursprung in unterschiedlichen Forschungsgebieten. Die Untersuchung von Korrelationen als Signatur einer Spezies hat ihren Ausgangspunkt in der Informationstheorie, die eine Symbolsequenz (z.B. eine DNA-Sequenz) als eine Nachricht mit syntaktischem Informationsgehalt auffasst. Die syntaktische Information bezieht sich dabei auf Wahrscheinlichkeiten für das Auftreten einzelner Symbole oder Symbolgruppen. Diese Wahrscheinlichkeiten spiegeln „Abhängigkeiten“ zwischen den Symbolen innerhalb der Sequenz wider, die im Folgenden als Korrelationen bezeichnet werden. Ein weiteres Forschungsgebiet ist die Theorie stochastischer Prozesse. Stochastische Prozesse können als Modell für die Implementierung solcher Korrelationen in Zeitreihen verwendet werden und haben in der Biologie u.a. zum Verständnis langreichweitiger Korrelation beigetragen (Li, 1989, 1991; Li et al., 1994). Um Korrelationen über größere Abstände effizient messen zu können, werden hier die Parameter eines diskreten autoregressiven Prozesses, der zur Modellierung von Symbolsequenzen mit Markov-Eigenschaft herangezogen werden kann, verwendet. Die auf diese Weise erhaltenen Korrelationsverläufe werden mit Methoden der Clusteranalyse untersucht, deren Ziel die Einteilung einer Menge unterschiedlicher Objekte in Gruppen mit gemeinsamen Eigenschaften ist. Hier kommen insbesondere bioinformatische Methoden zum Einsatz. Neben diesen etablierten Methoden werden im Folgenden auch Werkzeuge vorgestellt, die im Laufe der Arbeit neu entwickelt wurden.

Die Beschreibungen der verwendeten Methoden in diesem Kapitel repräsentieren genau die Ausschnitte aus den unterschiedlichen Forschungsgebieten, die zum Verständnis der in den folgenden Kapiteln erzielten Ergebnisse benötigt werden. Einige Aspekte des vorliegenden Kapitels entsprechen der Darstellung, wie wir sie in Hütt und Dehnert (2006) formuliert haben.

1.1 Stochastische Prozesse

Stochastische Prozesse sind Modelle für zeitlich geordnete, zufällige Vorgänge. Aus mathematischer Sicht ist ein stochastischer Prozess eine Folge von Zufallsvariablen, die einer zugrunde liegenden Verteilung gehorchen. Dabei nehmen die Zufallsvariablen X_t die Zustände i an, $X_t = i$, deren Wertebereich durch den Zustandsraum festgelegt ist. Eine Familie von Zufallsvariablen $\{X_t, t \in T\}$ mit Werten in Σ heißt dann stochastischer Prozess mit dem Parameterbereich (der Indexmenge) T und dem Zustandsraum Σ. Der Zustandsraum Σ sei hier stets abzählbar, d.h. endlich

oder abzählbar unendlich. Der Parameter t repräsentiert im Allgemeinen die Zeit. Für die hier diskutierten Anwendungen stellt dieser Parameter die Symbolnummer, also die Position eines Symbols (Zustands) entlang der Sequenz, dar. Im Falle $T = \mathbb{N} = \{0,1,2,3,\ldots\}$ bezeichnet man $\{X_t\}$ als diskreten stochastischen Prozess im Sinne eines Prozesses mit diskretem Parameter (Zeit). Die im Folgenden betrachteten Prozesse zeichnen sich dadurch aus, dass ihre statistischen Eigenschaften invariant gegenüber Verschiebungen der Zeit sind. Für einen solchen *stationären* Prozess $\{X_t\}$ gilt, dass für beliebige Zeitpunkte $t_1,\ldots,t_k$ und h aus der Indexmenge T die gemeinsame Verteilung von $\{X_{t_1},\ldots,X_{t_k}\}$ die gleiche ist wie für $\{X_{t_1+h},\ldots,X_{t_k+h}\}$. Die Wahrscheinlichkeiten hängen damit nicht vom Beobachtungszeitpunkt, also in unserem Fall der Position innerhalb der Sequenz, ab. Dies ist eine sehr starke Forderung an den Prozess, die für reale Daten schwer nachzuweisen ist und oft nicht erfüllt wird. Trotzdem können auch bei Verletzung der Stationaritätsannahme Prozesse mit geeigneten Methoden analysiert werden. Eine ausführliche Einführung in stochastische Prozesse gibt das Buch von Karlin und Taylor (1975).

1.1.1 Markov-Prozesse

In der einfachsten Form eines stochastischen Prozesses sind die Zufallsvariablen unabhängig. Sei $p(x_1,\ldots,x_n) = P(X_1 = x_1,\ldots,X_n = x_n)$ die Wahrscheinlichkeit für die Beobachtung des n-Wortes $x_1,\ldots,x_n$ mit x_i aus dem Zustandsraum Σ, dann gilt im Falle eines unabhängigen Prozesses[1]

$$p(x_1,\ldots,x_n) = p(x_1)\cdot\ldots\cdot p(x_n)\ . \tag{1.1}$$

Sind die Zufallsvariablen nicht unabhängig voneinander, so spiegelt sich diese Korrelation in den n-Wort Verteilungen der Symbolsequenzen wider. Bedingte Wahrscheinlichkeiten ermöglichen den Zugang zu diesen Abhängigkeiten. Die bedingte Wahrscheinlichkeit $p(x_n|x_{n-1},\ldots,x_1)$ beschreibt die Wahrscheinlichkeit für das Beobachten des Symbols x_n unter der Bedingung, dass die vorangegangenen Symbole x_1 bis x_{n-1} beobachtet wurden, und ist definiert als

$$p(x_n|x_{n-1},\ldots,x_1) = \frac{p(x_n,x_{n-1},\ldots,x_1)}{p(x_{n-1},\ldots,x_1)}\ . \tag{1.2}$$

Damit fragt die bedingte Wahrscheinlichkeit explizit nach einer Korrelation zwischen dem Symbol x_i und seinen Vorgängern in der Symbolfolge.

Ein stochastischer Prozess $\{X_n\}$ heißt Markov-Prozess (erster Ordnung), falls gilt:

$$p(x_n|x_{n-1},\ldots,x_1) = p(x_n|x_{n-1})\ . \tag{1.3}$$

Es hat also ausschließlich das unmittelbar vorangegangene Symbol x_{n-1} einen Einfluss auf die Wahrscheinlichkeitsverteilung des Symbols x_n. Ein homogener Markov-Prozess $\{X_n\}$ wird vollständig beschrieben durch den (diskreten und endlichen) Zustandsraum $\Sigma = \{a_1,a_2,\ldots,a_N\}$ der Größe N, eine Startverteilung p_0 und ein System von Übergangswahrscheinlichkeiten. Für einen Markov-Prozess erster Ordnung ist dieses System zweidimensional. Die Übergangswahrscheinlichkeiten bilden eine $(N\times N)$-Matrix, die Übergangsmatrix Π mit

[1] Auf die Unterscheidung zwischen der Zufallsvariablen X_t und der Realisierung i in der Form $P(X_t = i)$ für die Wahrscheinlichkeit des Ereignisses i wird im Folgenden nur zurückgegriffen, wenn dies einem besseren Verständnis dient. Andernfalls wird die verkürzte Notation $p(i) \equiv P(X_t = i)$ vorgezogen, die eine klare Darstellung unterstützt.

$$p_{ij} := p(x_n = j | x_{n-1} = i), \qquad i, j \in \Sigma . \tag{1.4}$$

Die bedingten Wahrscheinlichkeiten beschreiben die Wahrscheinlichkeit, vom Zustand i direkt in den Zustand j zu gelangen. Für die Übergangsmatrix $\Pi = (p_{ij})$ gilt:

$$p_{ij} \geq 0 ; \quad \sum_{j \in \Sigma} p_{ij} = 1 \quad \forall i \in \Sigma . \tag{1.5}$$

Die Startverteilung p_0 ordnet jedem Element des Zustandsraumes eine Wahrscheinlichkeit dafür zu, dass der Prozess mit diesem Element beginnt.

Eine direkte Verallgemeinerung des Markov-Prozesses erster Ordnung ergibt sich, wenn man die Abhängigkeit der bedingten Wahrscheinlichkeiten auf p vorangegangene Symbole ausdehnt. Ein stochastischer Prozess $\{X_n\}$ heißt Markov-Prozess der Ordnung p, falls gilt:

$$p(x_n | x_{n-1}, \ldots, x_1) = p(x_n | x_{n-1}, \ldots, x_{n-p}), \qquad n > p . \tag{1.6}$$

Man spricht bei einem solchen Prozess auch von einem Prozess mit einem „Gedächtnis der Länge p“ (Ebeling et al., 1998). Ein Markov-Prozess kann für den Fall $p > 1$ durch ein $(p+1)$-dimensionales System von Übergangswahrscheinlichkeiten (verallgemeinerte Matrix oder Tensor) und eine vorgegebene Liste mit den Wahrscheinlichkeiten aller p-Worte als Startverteilung angegeben werden.

Fasst man DNA-Sequenzen als Realisierung eines diskreten stochastischen Prozesses auf, so kann man die Parameter aus den Sequenzen extrahieren, d.h. schätzen und mit biologischen Eigenschaften in Verbindung bringen. Dieser Zugang zu den statistischen Besonderheiten einer DNA-Sequenz und ihren biologischen Ursachen wird in der vorliegenden Arbeit verfolgt.

1.2 Informationstheoretische Maße

Die in der Mitte des letzten Jahrhunderts begründete Informationstheorie hat das Ziel, den Begriff der *Information* rein statistisch zu erfassen. Einen großen Beitrag dazu lieferte der Mathematiker Claude E. Shannon mit seiner Arbeit *A mathematical theory of communication* (Shannon, 1948). In der Vorstellung der Informationstheorie ist eine (unendliche) Sequenz eine Realisierung eines stationären Prozesses. In diesem Prozess liegen die Wahrscheinlichkeiten als reale Parameter vor, und die Betrachtung der Sequenz erlaubt eine Schätzung dieser Parameter aus den beobachteten Häufigkeiten. Methoden der Informationstheorie extrahieren so aus beobachteten Sequenzen Eigenschaften des Prozesses. Betrachten wir die folgende Situation: Die Symbole i aus dem Zustandsraum Σ, dem „Alphabet“ der Sequenz, treten mit den Wahrscheinlichkeiten p_i auf, die in einer diskreten Verteilung $P = p_1, \ldots, p_N$ zusammengefasst werden. Dabei ist $N = |\Sigma|$ die Größe des Zustandsraums. Die Shannon-Entropie beschreibt nun den mittleren Informationsgewinn bei Beobachtung eines (statistischen) Ereignisses i aus dem Zustandsraum.[2] In einem axiomatischen Zugang stellt Shannon dabei drei essentielle Forderungen an das Informationsmaß der Entropie

[2] Statt als mittlerer Informationsgewinn kann die Entropie auch als die mittlere Unsicherheit bei der Vorhersage eines Ereignisses oder die mittlere Menge an Information betrachtet werden, die man benötigt, um ein Ereignis vorherzusagen.

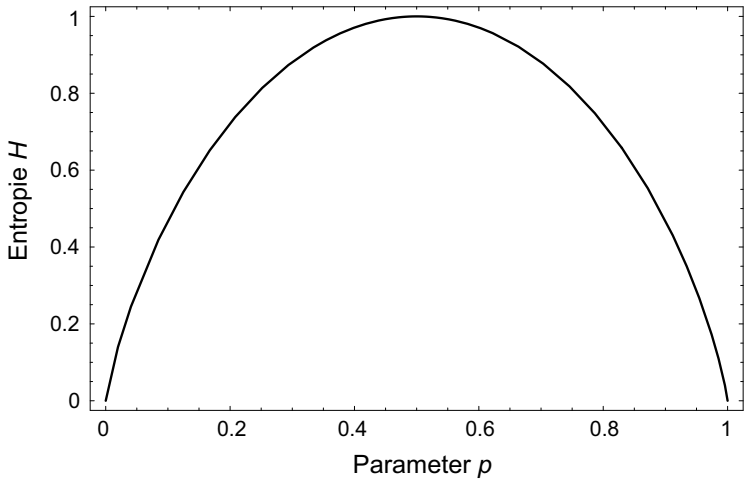

Abb. 1.1. Entropie H aus Gleichung (1.8) in Abhängigkeit von p für das Beispiel des Münzwurfexperiments. (Angepasst aus: Shannon (1948).)

H: Erstens Stetigkeit in p_i, zweitens, dass H maximal ist, wenn alle möglichen Ereignisse mit der gleichen Wahrscheinlichkeit eintreten, und drittens, dass für zusammengesetzte Ereignisse die Entropie H durch die gewichtete Summe der Entropien der Einzelereignisse beschrieben werden kann.

Es kann gezeigt werden, dass das folgende Maß, die Shannon-Entropie, diese Eigenschaften erfüllt:

$$H = -\sum_{i=1}^{N} p_i \log_\lambda p_i \,, \tag{1.7}$$

wobei λ als Basis des Logarithmus die Einheit von H festlegt. Wählt man λ gleich der Größe des Alphabets Σ, so lassen sich auf verschiedenen Alphabetgrößen basierende Entropien direkt miteinander vergleichen. Eine häufige Wahl ist $\lambda = 2$, denn damit ergibt sich die Entropie in Einheiten von einem Bit. Um ein Gefühl für die Entropie zu erhalten, betrachten wir ein Zufallsexperiment, bei dem eine (nicht notwendigerweise symmetrische) Münze geworfen wird. Die Wahrscheinlichkeit für das Ereignis *Kopf* ist durch p gegeben, die für *Zahl* durch $q = 1 - p$. Somit erhält man für die Entropie:

$$H = -(p \log_2 p + q \log_2 q) \,. \tag{1.8}$$

In Abbildung 1.1 ist der graphische Verlauf der Entropie in Abhängigkeit von p aufgetragen. Diese Abbildung ist bereits in Shannons Originalarbeit (Shannon, 1948) als erläuterndes Beispiel aufgeführt, um zwei der drei geforderten Eigenschaften der Entropie zu überprüfen, nämlich die Stetigkeit und das Erreichen des Maximums bei gleichwahrscheinlichen Ereignissen. Die Entropie H ist Null, wenn der Ausgang des Zufallsexperiments bekannt ist, d.h. wenn ein p_i gleich Eins ist. Für den Fall des Münzwurfs ist dies bei $p = 1$ und $p = 0$ (also $q = 1$) der Fall. In allen anderen Fällen ist H positiv. H ist maximal, wenn alle N möglichen Ereignisse mit der gleichen Wahrscheinlichkeit eintreten, also mit $1/N$. Für das Münzwurfexperiment bedeutet das $p = q =$

$1/2$. Dies ist auch intuitiv klar: Wenn die Chancen 50 zu 50 stehen, ist die Unsicherheit für das Registrieren jedes der beiden Ereignisse *Kopf* und *Zahl* maximal.

Eine Bemerkung ist an dieser Stelle angebracht: Stellt man sich die Wahrscheinlichkeiten p_i in Gleichung (1.7) durch relative Häufigkeiten approximiert vor, so wird durch die Entropie eine (lange) Sequenz in eine einzelne Zahl übersetzt. Eine wichtige Voraussetzung für eine Interpretierbarkeit des Ergebnisses (und damit für die Anwendbarkeit solcher Methoden) ist die Stationarität der Sequenz. Am Beispiel der Entropie kann man die Bedeutung der Stationarität für die Anwendung von statistischen Methoden illustrieren. Für eine sehr geordnete Sequenz auf einem binären Zustandsraum, die nach der Hälfte einmal den Zustand wechselt, also

$$11111\ldots1100000\ldots00\,,$$

würde sich aus der Sequenz $p_1 = p_0 = 0.5$ ergeben, was unsinnigerweise auf eine maximale Entropie führt. Die Ursache liegt in der Verletzung der Stationarität: Die Wahrscheinlichkeiten hängen vom Beobachtungszeitpunkt, also der Position innerhalb der Sequenz ab.

Entropien höherer Ordnung, die eine Verallgemeinerung der Shannon-Entropie darstellen, werden bei der Interpretation von Markov-Prozessen in Anhang A.1 diskutiert.

1.2.1 Transinformation

Neben der Entropie hat sich ein weiteres Maß der Informationstheorie als sehr nützlich bei der Beschreibung von Korrelationen in DNA-Sequenzen erwiesen. Die Transinformation (engl. *mutual information*) kann als Differenz von Shannon-Entropien dargestellt werden (Shannon, 1948; Herzel und Ebeling, 1985) und beschreibt für zwei Ereignisse, die sich gegenseitig beeinflussen, um wieviel die Unbestimmtheit des zweiten Ereignisses durch Kenntnis des ersten Ereignisses im Mittel kleiner wird.

Betrachtet man eine Symbolsequenz auf dem Alphabet Σ und bezeichnet mit $p^{(k)}(i,j)$ die Wahrscheinlichkeit, die Symbole i und j im Abstand k zu beobachten, und mit $p(i)$ und $p(j)$ die Einzelwahrscheinlichkeiten der entsprechenden Symbole, so ist die *Transinformation* $I(k)$ als Funktion von k definiert als

$$I(k) = \sum_{(i,j)\in\Sigma^2} p^{(k)}(i,j)\log_\lambda \frac{p^{(k)}(i,j)}{p(i)\,p(j)}\,. \tag{1.9}$$

Die Transinformation[3] hat einige Eigenschaften, die hier kurz angesprochen werden sollen, indem die Grenzwerte des Verhaltens von $I(k)$ betrachtet werden. Wir tun dies für den Spezialfall $k = 1$, also für benachbarte Symbole. Man hat dann Paarwahrscheinlichkeiten $p^{(1)}(i,j) \equiv p(i,j) \equiv p_{ij}$. Nehmen wir an, ein p_{ij} wäre Eins. Aus Symmetriegründen, da sonst p_{ji} nicht verschwinden könnte, muss i gleich j sein und damit folgt $p_i = 1$ und $I = 0$. Ein weiterer wichtiger Spezialfall ist der einer unabhängigen Abfolge von Zuständen. In diesem Fall ist $p_{ij} = p_i p_j$ und damit erneut $I = 0$, weil jeder Summand in Gleichung (1.9) den Faktor $\log_\lambda 1 = 0$ enthält. Im ersten Fall ist die

[3] Die in Gleichung (1.9) angegebene Größe $I(k)$ bezeichnet man oft auch als Transinformationsfunktion, da sie vom Abstand k der beiden Symbole abhängt. Die eigentliche Transinformation I ergibt sich dann aus dem Spezialfall direkt benachbarter Symbole, $I = I(1)$.

Sequenz maximal korreliert und damit vollständig bestimmt. Die Kenntnis eines Symbols liefert keine Information über das benachbarte Symbol. Im zweiten Fall, der vollkommen unkorrelierten Sequenz lässt die Kenntnis über ein Symbol keinen Rückschluss auf das benachbarte Symbol zu. Der Informationsgewinn ist also ebenfalls Null. Jede andere Wahrscheinlichkeitsverteilung führt auf eine nicht verschwindende Transinformation. Es ist gerade diese Eigenschaft, die trivialen Fälle von Paarkorrelationen (eine konstante Sequenz und eine vollkommen zufällige Sequenz) auf $I = 0$ abzubilden, die die Transinformation als Maß für Komplexität nahelegt (Ebeling et al., 1998). Betrachten wir nun wieder den allgemeinen Fall $I(k)$. Je mehr die Verteilung $p^{(k)}(i, j)$ im Mittel von der Produktform (also der unabhängigen Verteilung) $p(i)p(j)$ abweicht, umso größer ist der Wert der Transinformation. Die Menge an Information, die ein beliebiges Symbol über das k Positionen entfernte Symbol enthält, beschreibt im Wesentlichen die Stärke der Korrelation, also die Stärke des „Zusammenhangs“ zwischen zwei Symbolen im Abstand k. Die Transinformation $I(k)$ hängt nur von zwei Symbolen ab und ist somit schon bei einer geringeren Sequenzlänge verlässlich aus der Sequenz zu schätzen. In Herzel und Grosse (1995, 1997) wird die Transinformation mit anderen Korrelationsmaßen verglichen und Fehlerabschätzungen für $I(k)$ bei endlicher Sequenzlänge angegeben.

Hier soll anhand einer konstruierten Sequenz noch einmal die Funktionsweise der Transinformation erläutert werden. Dazu wird in einer zufälligen, also einer unabhängigen, Symbolsequenz ein variables Muster induziert und die so veränderte Sequenz mit der Transinformation analysiert. Um das Beispiel überschaubar zu gestalten, basiert die Sequenz auf einem binären Alphabet $A = \{0, 1\}$ mit einer Gleichverteilung der Wahrscheinlichkeiten $P(X = 0) = P(X = 1) = 0.5$. Aus der geforderten Unabhängigkeit der Symbole in der Sequenz folgt, dass die Wahrscheinlichkeit von zwei Symbolen im Abstand k gleich dem Produkt der Einzelwahrscheinlichkeiten ist und die Transinformation damit für alle Symbolabstände k gleich Null. In dieser Sequenz werden Segmente mit einer Länge von 10 Basen und einer internen Mutationsrate von 10% an zufällig ausgewählten Positionen eingefügt. Es liegt also innerhalb eines eingefügten Segments jedes Symbol in jeder Position mit einer Wahrscheinlichkeit von 90% vor. Die Wahrscheinlichkeit einer solchen Einfügung wird auf 5% gesetzt. In Abbildung 1.2 a ist der Aufbau der Sequenz graphisch visualisiert. Die Berechnung der Transinformation für diese Sequenz führt auf den in Abbildung 1.2 b dargestellten Verlauf. Die Konstruktion des variablen Sequenzabschnitts führt zu einer veränderten Wahrscheinlichkeitsverteilung für Symbole im Abstand k und somit auf mehrere Peaks in der Transinformationskurve. Die größte Amplitude zeigt der Verlauf für $k = 3$ als Folge einer deutlichen Korrelation in diesem Symbolabstand in der eingefügten Sequenz. Für Symbolabstände $k > 10$ ist die Transinformation Null.

Was hier von Hand eingeführt ist, tritt in DNA in vielfältiger Weise auf. Sequenzen, die Information tragen, unterscheiden sich von zufälligen Abfolgen von Symbolen genau dadurch, dass sie spezielle, Bestandteilen der Sequenz entsprechende Korrelationen aufweisen. Um diese Korrelationen geht es in der vorliegenden Arbeit.

1.2.2 DAR(p)-Prozesse

Eine andere, modellhaftere Art, die Stärke der Korrelation zweier Symbole im Abstand k zu quantifizieren, ist durch die Parameter eines diskreten autoregressiven Prozesses $p$ter Ordnung, eines DAR(p)-Prozesses, gegeben. Ein solcher kann zur Generierung einer Symbolsequenz mit einer

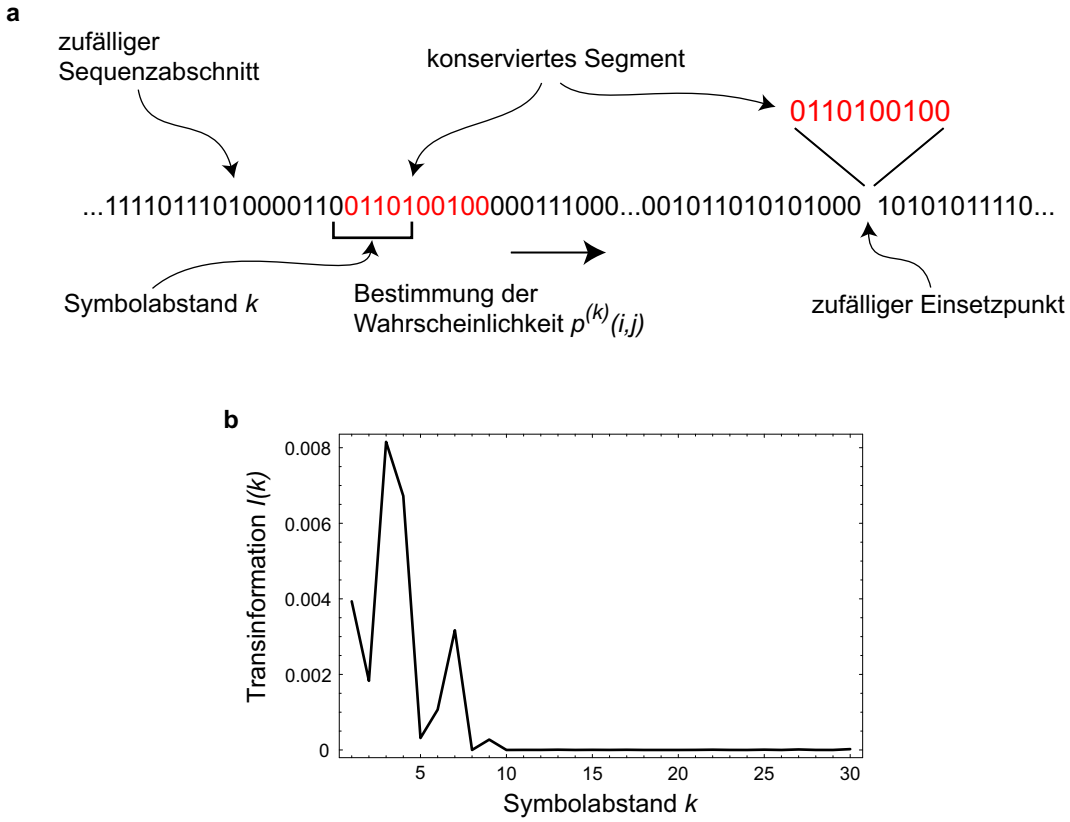

Abb. 1.2. **a** Konstruktionsprinzip einer binären Sequenz mit induziertem Muster. **b** Transinformation $I(k)$ im Symbolabstand k für eine nach dem in Abbildungsteil **a** dargestellten Schema generierte Symbolsequenz.

Markov-Eigenschaft der Ordnung p herangezogen werden, also einer Realisierung eines Markov-Prozesses in dem die Verteilung von X_n von $X_{n-1},\ldots,X_{n-p}$ abhängt. Er kann umgekehrt, wie später dargestellt werden soll, auch zur Messung der Korrelationen verwendet werden. Da eine Vielzahl der Ergebnisse in der vorliegenden Arbeit mit Hilfe dieser Beschreibung von Korrelationen gewonnen wurden, wird der Prozess hier im Detail diskutiert. Der Prozess wird bestimmt durch eine stationäre Marginalverteilung von X_n und mehreren anderen Parametern, welche unabhängig von der Randverteilung die Korrelationsstruktur bestimmen. Die Kernidee einer solchen Sequenzerzeugung ist dabei eine Rekursion. Die ersten p Symbole einer zu erzeugenden Sequenz sind gegeben (gezogen aus dem Alphabet nach einer gegebenen Startverteilung), und man bestimmt nun das $(p+1)$te Symbol entweder durch Rückgriff auf eines der vorangegangenen Symbole oder durch erneute zufällige Wahl aus dem Alphabet. Die Parameter des Prozesses legen die Wahrscheinlichkeit für ein Zurückgreifen und ein zufälliges Auswählen fest. Nach dem $(p+1)$ten Symbol bestimmt man nun das $(p+2)$te Symbol und so fort.

Sei X_n das $n$te Symbol in einer durch einen DAR(p)-Prozess generierten Sequenz. Dann ist X_n gegeben durch die folgende rekursive Anweisung (Jacobs und Lewis, 1978):

$$X_n = V_n X_{n-A_n} + (1-V_n)Y_n\,, \qquad n = p,\, p+1,\, p+2,\,\ldots\,. \tag{1.10}$$

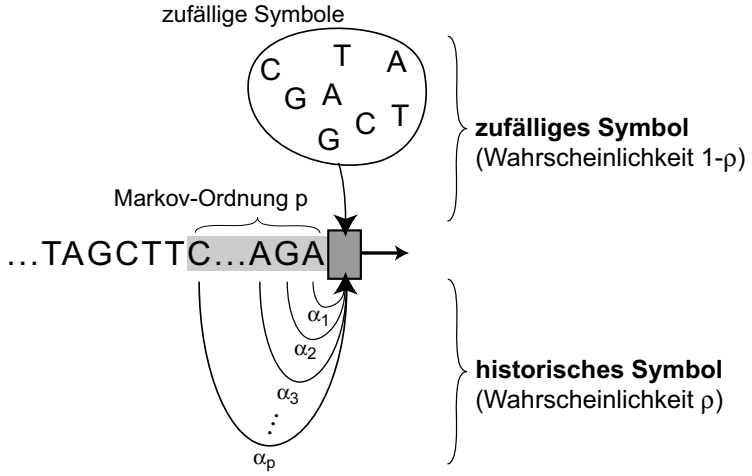

Abb. 1.3. Schematische Darstellung des DAR(p)-Prozesses aus Gleichung (1.10). Ein neues Symbol (Kasten am rechten Sequenzende) wird der Sequenz entweder durch Ziehen eines zufälligen Symbols (oberer Bildteil; Wahrscheinlichkeit $1-\rho$) oder durch Rückgriff auf ein Vorgängersymbol (unterer Bildteil; Wahrscheinlichkeit ρ) bestimmt. In diesem unteren Zweig geschieht mit der Wahrscheinlichkeit α_i ein Rückgriff um i Stellen. Die maximale Rückgriffweite ist durch die festgelegte Markov-Ordnung p gegeben. (Aus: Hütt und Dehnert (2006).)

Der erste Term in diesem rekursiven Modell ist für die Markov-Eigenschaft verantwortlich, während der zweite Term unkorrelierte, zufällig gezogene Symbole aus dem Alphabet in die Sequenz einfließen lässt. Die Zufallsvariable V_n nimmt die Werte 0 und 1 an und wirkt damit als Schalter zwischen den zwei Termen der rechten Seite von Gleichung (1.10). Der Wert $V_n = 1$ tritt mit der Wahrscheinlichkeit ρ ein, der Wert $V_n = 0$ mit der verbleibenden Wahrscheinlichkeit $1-\rho$. Die weiteren Parameter dieses Prozesses verbergen sich in der Zufallsvariablen A_n. Diese nimmt die Werte $1, 2, \ldots, p$ an, und zwar mit den Wahrscheinlichkeiten $\alpha_1, \alpha_2, \ldots, \alpha_p$. Die Werte α_k regulieren dabei, wie oft das Symbol X_n in der Sequenz durch das Symbol X_{n-k}, das k Schritte in der Sequenz zurückliegt, determiniert wird, falls ein Rückgriff erfolgt. Als letzten Baustein besitzt der Prozess die Zufallsvariable Y_n die Werte des Alphabets nach einer festzulegenden Verteilung π, der Marginalverteilung, annimmt. Die Zufallsvariablen V_n, A_n und Y_n werden als unabhängig angesehen. Die Sequenz X_n hat eine Markov-Eigenschaft $p$ter Ordnung, wobei die Werte α_k per Konstruktion die Stärke der Korrelation im Abstand k beschreiben. Abbildung 1.3 fasst die Funktionsweise dieser rekursiven, durch den DAR(p)-Prozess gegebenen Erzeugung einer Symbolsequenz schematisch zusammen. Eine rein formale Definition der DAR(p)-Prozesse und einige mathematische Eigenschaften werden im Anhang A diskutiert.

Der DAR(p)-Prozess stellt eine sehr parametereffiziente Weise dar, um einen Markov-Prozess zu realisieren. Ein solcher DAR(p)-Prozess ist ein Spezialfall eines Markov-Prozesses und beschreibt nicht alle möglichen Markov-Prozesse $p$ter Ordnung. Es gibt eine Vielzahl von Prozessen, die Zahlenfolgen mit einer Markov-Eigenschaft erzeugen. Das Bemerkenswerte an Gleichung (1.10) ist, dass hier eine Sequenz auf einem beliebigen Zustandsraum generiert werden kann. So ist zum Beispiel nicht erforderlich, dass es Abstände (bzw. die Möglichkeit, Elemente des Zustandsraums

zu addieren oder subtrahieren) auf dem Zustandsraum gibt. Daher kann man diesen Prozess besonders gut nutzen, um Symbolsequenzen zu erzeugen.

Neben der Simulation von Symbolsequenzen mit festlegbaren Parameterkonstellationen können alle Parameter auch aus einer gegebenen Sequenz geschätzt werden. Wir werden sehen, dass der Parametervektor $\vec{\alpha}$ die Korrelationsstärke im Abstand k in sehr guter Weise quantifiziert. Der Schätzprozess besteht aus zwei Schritten. Im ersten Schritt wird die Korrelationsstärke mit Hilfe einer empirischen Autokorrelationsfunktion bestimmt. Dieser *ad hoc* Schätzer wurde in Zusammenhang mit dem DAR(p)-Prozess eingeführt, ist aber in seiner Berechnung nicht von einem solchen Prozess abhängig. Der *ad hoc* Schätzer $\hat{r}(k)$ ist für die Korrelation zweier Symbole im Abstand k wie folgt definiert (Jacobs und Lewis, 1983):

$$\hat{r}(k) = 1 - \sum_{a_i \in A} B_m(k, a_i) \frac{1}{1 - \pi(a_i)} \,, \tag{1.11}$$

mit $k \in \mathbb{N}$ und der Marginalverteilung π und

$$B_m(k, a_i) = \frac{1}{m-k} \sum_{a_j \neq a_i \in A} \sum_{l=1}^{m-k} \delta_{a_j}(x_l) \delta_{a_i}(x_{l+k}) \,, \tag{1.12}$$

wobei die Indikator-Funktion $\delta_y(x) = 1$ für $x = y$ und $\delta_y(x) = 0$ für $x \neq y$ ist. Eine aus algorithmischer Sicht effiziente Umsetzung des Schätzprozesses ist in Hameister (2006) beschrieben.

Der zweite Schritt führt von den Größen $\hat{r}(k)$ zu den tatsächlichen Parametern des DAR(p)-Prozesses. Um den Parametervektor α zu erhalten, muss ein nichtlineares Gleichungssystem gelöst werden, welches die (theoretischen) r und α-Parameter in Verbindung setzt (Jacobs und Lewis, 1978). Dieses als Yule-Walker-Gleichungen bezeichnete System kann mit Hilfe der Umformung $\phi_k := \rho\alpha_k$ in ein lineares Gleichungssystem der Form $Ax = b$ in ϕ_k überführt werden, was bei der Lösung zu einem erheblichen Rechenzeitgewinn führt und darüber hinaus erlaubt, die Eindeutigkeit der Lösung an der Determinanten der Matrix A abzulesen. Nach der Transformation lauten die Yule-Walker-Gleichungen damit:

$$\begin{array}{rcccccccc} r(1) & = & \phi_1 r(0) & + & \phi_2 r(1) & + & \ldots & + & \phi_p r(p-1) \,, \\ r(2) & = & \phi_1 r(1) & + & \phi_2 r(0) & + & \ldots & + & \phi_p r(p-2) \,, \\ \vdots & & \vdots & & \vdots & & & & \vdots \\ r(p) & = & \phi_1 r(p-1) & + & \phi_2 r(p-2) & + & \ldots & + & \phi_p r(0) \,, \end{array} \tag{1.13}$$

mit $r(0) = 1$.

Der Parametervektor $\vec{\alpha}$ erfüllt die Normierung $\sum_{k=1}^{p} \alpha_k = 1$ und somit gilt für den Parameter ρ

$$\sum_{k=1}^{p} \phi_k = \sum_{k=1}^{p} \alpha_k \rho = \rho \sum_{k=1}^{p} \alpha_k = \rho \,. \tag{1.14}$$

Durch Einsetzen von $\hat{r}(1), \hat{r}(2), \ldots, \hat{r}(p)$ für $r(1), r(2), \ldots, r(p)$, können die p Gleichungen für die p Parameter mit $\phi_k = \rho\alpha_k$ und $k = 1, \ldots, p$ gelöst werden. Der Vektor $\vec{\alpha}$, den man durch diesen Schätzprozess erhält, wird im Folgenden als die Markov-Repräsentation der Korrelationsstärke

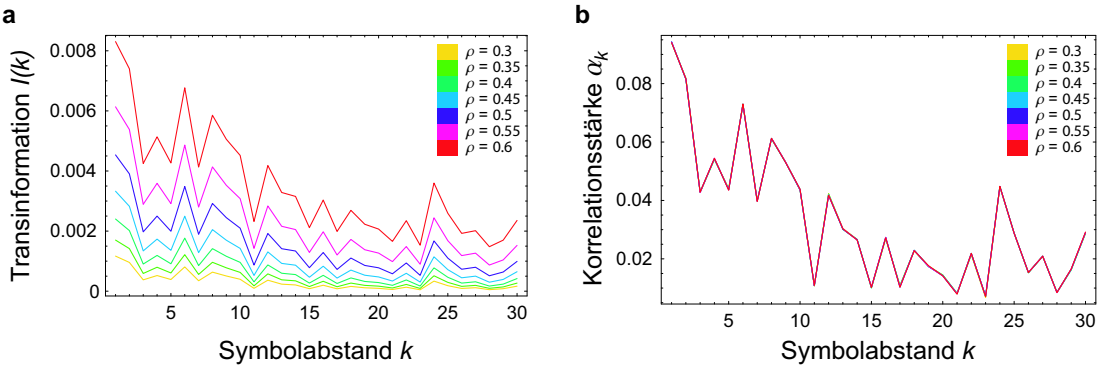

Abb. 1.4. **a** Transinformation $I(k)$ im Symbolabstand k für sieben Realisierungen eines DAR(30)-Prozesses mit $\rho = 0.3, 0.35, 0.4, \ldots, 0.6$ bei identischer Wahl des Parametervektors $\vec{\alpha}$ und Marginalverteilung π. **b** Schätzung der Korrelationsstärke in Form der α-Vektoren für die der Berechnung der Transinformation $I(k)$ in Teilabbildung **a** zugrundeliegenden Realisierungen des DAR(30)-Prozesses. Die Markov-Repräsentation in **b** führt für alle sieben Sequenzen auf nahezu identische Korrelationskurven.

bezeichnet. In der Interpretation als Parameter $\vec{\alpha}$ des in Gleichung (1.10) angegebenen Prozesses beschreibt jede Komponente α_k dieses Vektors die Wahrscheinlichkeit für den Rückgriff um k Stellen in der Sequenz bei der Bestimmung des neuen Symbols. Prinzipiell kann die Bestimmung der Korrelationsstärke der Markov-Repräsentation zu einzelnen negativen Komponenten im Vektor $\vec{\alpha}$ führen. Die Interpretation der α_k als Parameter des DAR(p)-Prozesses geht damit verloren, da diese Wahrscheinlichkeiten darstellen und somit nicht negativ sein können. Hier wären Nebenbedingungen erforderlich, um eine Interpretation als Wahrscheinlichkeiten wieder herzustellen. Der große Vorteil dieses Korrelationsmaßes gegenüber der Transinformation $I(k)$ ist, dass der Schätzprozess mit der Variablen ρ explizit die Menge an zufälliger Sequenz (also an „Hintergrundrauschen“) quantifiziert und dieser Beitrag nicht in der Korrelationsstärke beinhaltet ist.

Um dies zu illustrieren, betrachten wir eine Familie von Symbolsequenzen, die mit Hilfe eines DAR(30)-Prozesses generiert worden ist. Der Parametervektor $\vec{\alpha}$ und die Marginalverteilung π sind für alle Sequenzen identisch. Ausschließlich der Parameter ρ, der den Anteil der Stochastizität determiniert, wird variiert. Die Berechnung der Transinformation führt zu den Korrelationskurven in Abbildung 1.4 a. Es ist deutlich zu sehen, dass alle Kurven die gleiche Struktur aufweisen, aber horizontal zueinander verschoben sind. Die Symbolsequenz, deren Korrelationskurve am oberen Ende der Kurvenschar liegt, zeigt die größten Werte in der Transinformation, da diese Sequenz im Vergleich zu den anderen Sequenzen die geringste Stochastizität aufweist. Die Korrelationen innerhalb der Sequenz werden also deutlicher in der Transinformation abgebildet. Umgekehrt verfügt die Sequenz, die zur untersten Korrelationskurve in der Schar führt, über einen hohen Rauschanteil in Form von zufälligen Sequenzabschnitten. Schätzt man die Parameter eines DAR(30)-Prozesses aus diesen Symbolsequenzen, so ergeben die α-Vektoren eine Schar von Korrelationskurven, die nahezu übereinander liegen, wie in Abbildung 1.4 b deutlich wird. Im Gegensatz zur Transinformation führt der unterschiedliche Anteil von Zufälligkeit hier nicht zu einer Verschiebung in Richtung der Ordinate. Diese Eigenschaft stellt den wichtigsten Vorteil der Markov-Repräsentation gegenüber der Transinformation als Korrelationsmaß dar. Neben ihrer Fähigkeit als Rauschfilter unterscheiden sich die Transinformation und die Markov-Repräsentation

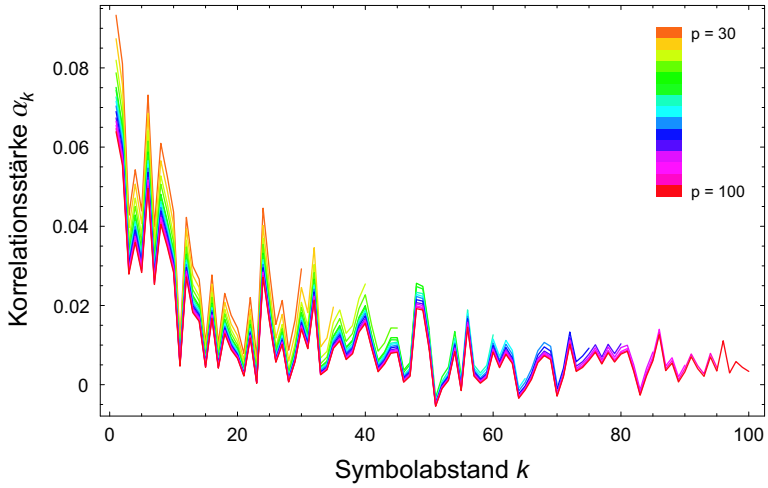

Abb. 1.5. Korrelationskurven der Markov-Repräsentation für das menschliche Chromosom 22 für unterschiedliche Markov-Ordnungen p. Beginnend mit $p = 30$ wird die Markov-Ordnung in Schrittweiten der Länge 5 sukzessiv erhöht bis zum Erreichen der Ordnung $p = 100$.

in einem weiteren Punkt. Dafür vergleichen wir die Korrelationskurve des simulierten „Chromosoms“ am oberen Ende der Kurvenschar in Abbildung 1.4 a mit den überlagerten Korrelationskurven in Abbildung 1.4 b. Die Ähnlichkeit zwischen diesen beiden Repräsentationen ist klar zu sehen, man erkennt aber auch Unterschiede. Es scheint, als würde die Markov-Repräsentation das deutlichere „Signal“ tragen (größere Peaks, vor allem bei größerem k). Dieser Eindruck wird sich im Folgenden bestätigen: Die Transinformation zeigt sich in der Anwendung zur Speziesunterscheidung als weniger geeignet als die Parameter eines DAR(p)-Prozesses. Die Transinformation ist jedoch das etabliertere Maß, wie eine große Zahl von auf der Transinformation basierenden Forschungsbeiträgen zu Korrelationen in DNA-Sequenzen belegen (siehe z.B. Herzel und Grosse (1995, 1997); Grosse et al. (2000); Holste et al. (2003); Li und Holste (2005)). Sie dient uns daher im Folgenden als Referenz.

Die Korrelationsstärke im Symbolabstand k ist im Fall des Parametervektors $\vec{\alpha}$ nicht unabhängig von dem Bereich von $k = 1, \ldots, p$. Der Grund dafür liegt in der Verrechnung der empirischen Autokorrelation durch die Yule-Walker-Gleichungen zu den Parametern des DAR(p)-Prozesses. Diese Normierungsfrage wird nun am Beispiel einer realen Sequenz, dem Chromosom 22 des Menschen, explizit vorgeführt. Um die Auswirkung dieser p-Abhängigkeit zu untersuchen, wird die Korrelationsstärke in Form der α-Vektoren für unterschiedliche p für das Chromosom berechnet. Der Parametervektor $\vec{\alpha}$ ist, unabhängig von p, durch die Yule-Walker-Gleichungen immer auf die Summe Eins normiert. Abbildung 1.5 zeigt die Markov-Repräsentation für unterschiedliche p. Man sieht, dass unabhängig von p die Korrelationskurven qualitativ alle den gleichen Verlauf zeigen. Bedingt durch die jeweilige Normierung der Kurven auf Eins müssen die Korrelationskurven für kleines p über denen für großes p liegen. Die Eigenschaften der Kurve bleiben bei Erweiterung der Markov-Ordnung jedoch im Wesentlichen erhalten, es gibt keinen „Shift“ von Korrelationsstärke von einer Region zur anderen. Alle Unterschiede skalieren damit linear mit der Normierung.

1.3 Clusteranalyse

Die Clusteranalyse ist ein Verfahren aus der multivariaten Statistik, mit deren Hilfe auf der Basis von Ähnlichkeiten oder Distanzen Objekte zu Clustern (Gruppen) mit gemeinsamen Eigenschaften zusammengefasst werden können. Clustermethoden werden seit den 1960er Jahren in der Biologie zur Konstruktion phylogenetischer Bäume eingesetzt (siehe z.B. Sokal und Sneath (1963); Saitou und Nei (1987); Nei und Kumar (2000)) und finden heute in zahlreichen wissenschaftlichen Disziplinen ihre Anwendung. Basierend auf einer Distanzmatrix, die den Abstand zwischen den Elementen der Analyse darstellt, wird mit einem Clusteralgorithmus diese Distanzmatrix in einen Baum übersetzt. Diese Elemente entsprechen hier den Korrelationskurven. Mit Hilfe von Bootstrap-Methoden kann untersucht werden, wie robust ein solcher Baum ist, und damit auch, wie aussagekräftig er ist.

1.3.1 Distanzmaße

Die Korrelationsstärken im Symbolabstand k, also die durch die Werteabfolge $\vec{\alpha} = \alpha_1, \ldots, \alpha_p$ gegebene Korrelationskurve einer DNA-Sequenz, bilden den Ausgangspunkt für die weiteren Analysen. In dieser Arbeit werden als DNA-Sequenzen typischerweise ganze Chromosomen eines eukaryotischen Genoms untersucht. Wir können daher von der Korrelationskurve eines Chromosoms sprechen. Der nächste Schritt besteht darin, ein Maß für die Unterschiedlichkeit der Korrelationskurven zweier Chromosomen zu definieren. Eine solche Distanz zwischen zwei höherdimensionalen Elementen kann auf unterschiedliche Weise definiert werden. Als sehr robust hat sich die L_1-Norm erwiesen, die als die Summe über die Beträge der Differenzen zweier Vektoren definiert ist (Kaufman und Rousseeuw, 1990):

$$d_{ij} = \left\| \vec{\alpha}^{(i)} - \vec{\alpha}^{(j)} \right\|_1 = \sum_{k=1}^{p} \left| \alpha_k^{(i)} - \alpha_k^{(j)} \right| , \tag{1.15}$$

wobei $\vec{\alpha}^{(s)} = (\alpha_1^{(s)}, \ldots, \alpha_p^{(s)})$ die Korrelationskurve des Chromosoms s bezeichnet und $\|\cdot\|_1$ die L_1-Metrik. In Abbildung 1.6 ist der Abstand zwischen zwei Korrelationskurven visualisiert. Die L_1-Metrik wird auch als „Manhattan“- oder „City Block“-Metrik bezeichnet. Durch die Berechnung aller möglichen paarweisen Abstände der Korrelationskurven erhält man eine Distanzmatrix, deren Hauptdiagonale mit Nullen besetzt ist.

1.3.2 Clusteralgorithmen

Es existiert eine große Anzahl von verschiedenen Clusteralgorithmen, die jeweils für verschiedene Bereiche und für spezifische Problemstellungen entwickelt worden sind. Die Unterschiede bestehen in den Annahmen zu den Eigenschaften der Daten und der Berechnung der Abstände zweier Cluster. Für einen Einblick in unterschiedliche Methoden der Clusteranalyse sei auf das Buch von Kaufman und Rousseeuw (1990) verwiesen. In der Biologie findet man Clusteralgorithmen vor allem bei einer Interpretation von Merkmalsähnlichkeiten in einem evolutionären Sinne, aber auch als rein statistisches Werkzeug zum Auffinden von Strukturen in numerischen Daten.

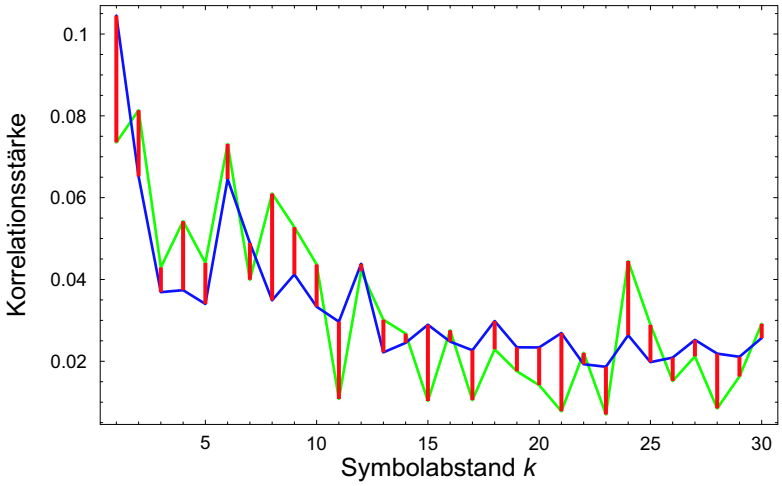

Abb. 1.6. Visualisierung der L_1-Distanz zwischen zwei Korrelationskurven der Ordnung $p = 30$. Die roten Balken zeigen die Abstände zwischen den Kurven in jedem Punkt, deren Summe den Abstand der Korrelationskurven beschreibt.

Ein Standardverfahren stellt der UPGMA-Algorithmus dar. Die Abkürzung UPGMA steht für *unweighted pair group method using arithmetic averages*. Am Anfang der Methode steht eine Erweiterung des Abstandsbegriffs. Der Abstand d_{ij} existiert bisher nur auf der Ebene von Objekten als Eintrag einer Distanzmatrix. Im Prozess der Baumkonstruktion werden Objektgruppen (Cluster) zusammengefasst, so dass ein allgemeinerer Abstandsbegriff erforderlich wird, der den Abstand zweier solcher Cluster angibt. Sei das Cluster C_k die Vereinigung zweier Cluster C_i und C_j, $C_k = C_i \cup C_j$, dann erhält man für den Abstand d_{kl} zwischen diesem neuen Cluster C_k und jedem anderen Cluster C_l den folgenden Ausdruck:

$$d_{kl} = \frac{d_{il} \cdot |C_i| + d_{jl} \cdot |C_j|}{|C_i| + |C_j|} , \tag{1.16}$$

wobei der Nenner gerade der Größe des neuen Clusters entspricht,

$$|C_i| + |C_j| = |C_k| . \tag{1.17}$$

Diese Gleichung ist das Kernstück des UPGMA-Algorithmus. Iterativ kann man nun, beginnend mit einer Menge von Clustern die jeweils nur ein Element enthalten, zu einer Baumstruktur gelangen, indem Cluster mit minimalem Abstand zusammengefasst werden und die Distanzmatrix mit Hilfe von Gleichung (1.16) neu bestimmt wird.

Der UPGMA-Algorithmus soll hier anhand eines einfachen Beispiels illustriert werden. Betrachten wir dazu die Korrelationskurven für das erste Chromosom des Menschen (HU1), der Maus (MU1) und der Ratte (RA1) in Abbildung 1.7 und die zugehörige nach Gleichung (1.15) bestimmte Distanzmatrix. Der erste Schritt besteht darin, jedem Element in der Analyse ein Cluster zuzuweisen: $C_1 = \text{HU1}$, $C_2 = \text{MU1}$ und $C_3 = \text{RA1}$. Die neue Nomenklatur führt auf die folgende Distanzmatrix:

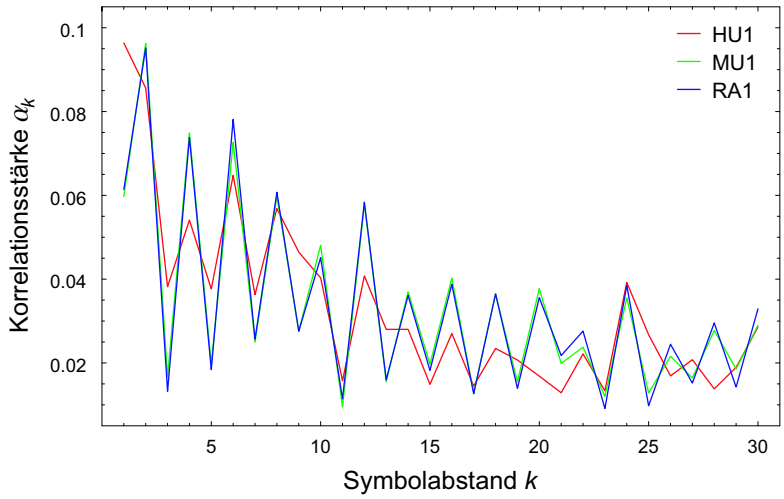

Abb. 1.7. Korrelationskurven für das erste Chromosom des Menschen (HU1), der Maus (MU1) und der Ratte (RA1).

	C_1	C_2	C_3
C_1	0	0.31	0.33
C_2	0.31	0	0.06
C_3	0.33	0.06	0

Die Ausgangssituation spiegelt sich in der graphischen Repräsentation wider, indem alle so initialisierten Cluster als Punkte auf der Höhe Null dargestellt werden (Abbildung 1.8 a). Nun werden die Cluster bestimmt, für die der Abstand in der Distanzmatrix minimal ist. Dieses Kriterium wird von den Clustern C_2 und C_3 erfüllt mit dem Abstand $d_{23} = 0.06$. Der Vereinigung dieser Cluster in $C_4 = C_2 \cup C_3$ folgt die Berechnung des Abstands des neuen Clusters C_4 zu allen verbleibenden Clustern. Man erhält den Abstand von C_4 zu C_1 durch

$$d_{41} = \frac{0.31 \cdot 1 + 0.33 \cdot 1}{2} = 0.32\,. \tag{1.18}$$

Als Nächstes werden C_2 und C_3 eliminiert, und C_4 wird der Clusterliste hinzugefügt. Damit erhält man die neue Distanzmatrix als:

	C_4	C_1
C_4	0	0.32
C_1	0.32	0

Auf der graphischen Ebene repräsentiert das Cluster C_4 einen neuen Knoten im Baum, der die Tochterknoten C_2 und C_3 hat. Die Astlängen der Cluster C_2 und C_3 sind jeweils $d_{23}/2 = 0.03$. Diese Konstruktion ist in Abbildung 1.8 b eingezeichnet. Die Anzahl der verbliebenen Cluster ist nun zwei. An dieser Stelle folgt algorithmisch der Schritt der Terminierung. Die Wurzel des Baums wird auf der Höhe $d_{41}/2 = 0.16$ in der graphischen Darstellung angebracht (Abbildung 1.8 c).

In Abhängigkeit von der Wahl der Distanzfunktion und des Clusteralgorithmus ändert sich natürlich das Ergebnis einer solchen Analyse. Im Rahmen von zwei Diplomarbeiten (Plaumann,

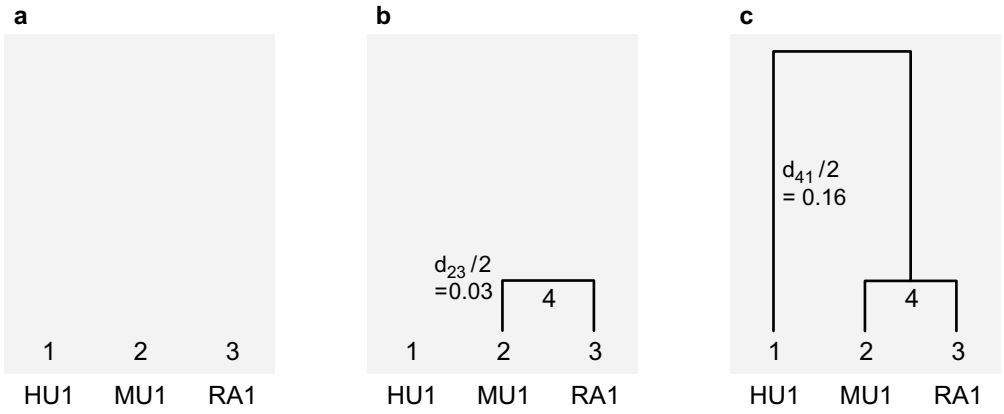

Abb. 1.8. Ergebnis für das Zahlenbeispiel zum UPGMA-Algorithmus. **a** zeigt den ersten Schritt der Baumkonstruktion, **c** stellt den Gesamtbaum dar. (In Anlehnung an: Hütt und Dehnert (2006).)

2003; Krauss, 2006) wurde eine Vielzahl von Distanzmaßen und Clusteralgorithmen an der hier beschriebenen Problematik angewandt. Dabei stellte sich heraus, dass für den größten Teil der betrachteten Metriken und Clusteralgorithmen die Ergebnisse von ähnlicher Qualität sind.

1.3.3 Bootstrap

Der nächste Schritt nach einer Baumkonstruktion ist die Bewertung des Baums in Bezug auf seine Robustheit gegenüber einer Variation der Daten. Die Schwierigkeit, die es dabei zu lösen gilt, ist, dass in der Praxis oft keine weiteren Daten zur Validierung der Ergebnisse zur Verfügung stehen. Um trotzdem zu einer Aussage zu gelangen, nutzt man *Bootstrap*-Methoden. Die Kernidee dabei ist, aus den Originaldaten eine Anzahl modifizierter Datensätze zu erzeugen, sogenannte Bootstrap-Samples, und mit denselben Methoden auszuwerten.[4] Die Modifikationen des Originaldatensatzes können etwa das Weglassen einzelner Segmente sein oder ein zufälliges *resampling*. Welche Art der Modifikation der Daten zielführend ist, hängt ganz entscheidend von den Daten und der darin vermuteten Information ab. Ziel der Bootstrap-Methoden ist es, die als nicht informationstragend angesehenen Strukturen in den Daten zu variieren oder zu manipulieren, um so zu überprüfen, ob die erzielten Ergebnisse von stochastischen Effekten beeinflusst werden. Für eine ausführliche Diskussion solcher Erzeugungsvorschriften und auch verschiedener Varianten des Bootstrap-Verfahrens sei auf das Buch von Efron und Tibshirani (1993) verwiesen. Bei der Bewertung von Clusterbäumen wird untersucht, wie häufig ein bestimmtes Cluster in einem Baum auftritt bei einer Modifizierung der der Analyse zugrunde liegenden Daten. Solche Bootstrap-Analysen liefern wichtige Indikatoren für die Robustheit einer Clusteranalyse. Bootstrap-Werte lassen sich für jeden internen Zweig (bzw. je nach Betrachtung: für jeden internen Knoten) ermitteln. Man zählt dabei nach, wie häufig in den Bäumen zu den modifizierten Datensätzen ein bestimmter

[4] In der klassischen Statistik werden Bootstrap-Verfahren eingesetzt, um Fehlerabschätzungen und Konfidenzintervalle zu erhalten. Im Fall, dass keine Annahme über die Verteilung der beobachteten Daten gemacht werden kann, ist diese Methode von großer Bedeutung. Basierend auf den Bootstrap-Samples bestimmt man eine empirische Verteilung der Statistik ohne Verteilungsannahme. Die resultierende empirische Verteilung wird dann zur Konstruktion des Konfidenzintervalls und weiterer Kenngrößen genutzt.

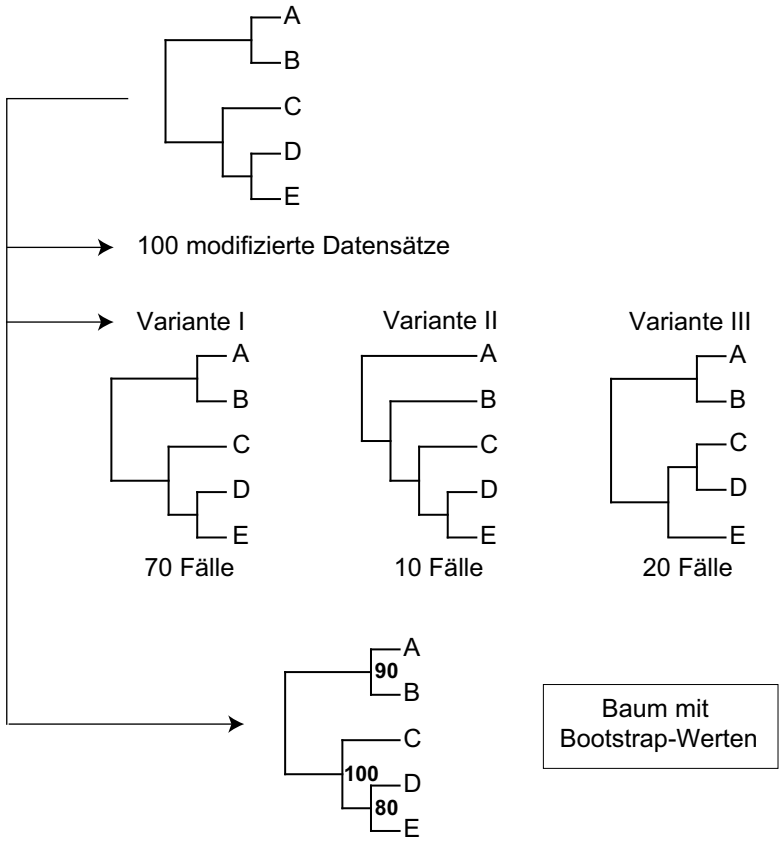

Abb. 1.9. Schematische Darstellung zu Bootstrap-Werten. Von einem Clusterbaum (oberes Bildelement) werden 100 Bootstrap-Replikate auf der Grundlage modifizierter Datensätze erzeugt. In diesem Gedankenexperiment bilden diese Replikate drei topologisch unterschiedliche Varianten (mittleres Bildelement). Der Bootstrap-Wert am Knoten gibt an, wie oft die Gruppe bestehend aus den Elementen rechts des Knotens in den Bäumen auftritt (unteres Bildelement). (In Anlehnung an: Hütt und Dehnert (2006).)

Zweig oder ein bestimmter interner Knoten vorkommt. Diesen Zahlenwert (typischerweise in Prozent, also bezogen auf 100 modifizierte Datensätze) schreibt man in dem Originalbaum an den entsprechenden Zweig oder Knoten. Diese Zahl gibt die *Bootstrap-Wahrscheinlichkeit* oder den Bootstrap-Wert eines Ergebnissegments an. Zweige in einem Clusterbaum mit sehr geringen Bootstrap-Werten lassen sich durch kleine Modifikationen an den zugrunde liegenden Daten aus dem Baum eliminieren. Abbildung 1.9 führt eine fiktive Bootstrap-Analyse vor.

Die Clusteranalyse basiert in dem hier betrachteten Fall auf den Korrelationsvektoren. Eine angemessene Form des Bootstrapping erhält man durch das zufällige Löschen paarweiser Komponenten $(\alpha_k^{(i)}, \alpha_k^{(j)})$ der Korrelationsvektoren $\vec{\alpha}^{(i)}$ und $\vec{\alpha}^{(j)}$ bei der Berechnung des Abstands d_{ij}. Die im Rahmen dieser Arbeit diskutierten Bootstrap-Bäume basieren auf 100 Distanzmatrizen, bei deren Erstellung zufällig jeweils 20% der paarweisen Komponenten in den Korrelationsvektoren gelöscht wurden. Zur Berechnung des *Consensus*-Baums wird das Programm *Consensus* mit der Option *50% majority-rule (extended)* des Software-Pakets PHYLIP eingesetzt. Eine weitere Art

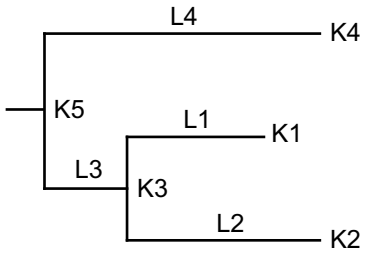

Abb. 1.10. Beispiel eines einfachen Cluster-Baums mit Bezeichnungen für die Blätter (K_1, K_2 und K_4), internen Knoten (K_3 und K_5) und Zweiglängen (L_1, L_2, L_3 und L_4). (Aus: Hütt und Dehnert (2006).)

von Bootstrap-Methoden basiert auf der Analyse von Teildatenmengen. Dieses Verfahren wenden wir in einer auf diese Fragestellung angepassten Vorgehensweise an, indem die Abhängigkeit der Korrelationskurven in Bezug auf die Sequenzlänge untersucht wird.

1.4 Tree Color Coding

Eine weitere Frage ist, wie interne Parameter der Analyse sich auf die Clusterbildung der Elemente im Baum auswirken. Die Schwierigkeit liegt darin, eine geeignete Darstellungsart zu finden, in der die Clusterbäume quasikontinuierlich in Abhängigkeit eines solchen Parameters beobachtet werden können. Die Darstellung als Baum mit Bootstrap-Werten eignet sich dafür nicht, da dort die Reihenfolge der Bäume nicht berücksichtigt wird. Hiermit könnte man höchstens überprüfen, welche Regionen des Baums besonders robust gegenüber einer Parameteränderung sind. Um eine solche Abhängigkeit der Clusterung zu untersuchen ist im Rahmen dieser Arbeit die Methode des *Tree Color Coding* (TCC) entwickelt worden. Dazu werden die Elemente des Baums in eine universelle Reihenfolge gebracht, es wird jedem Chromosom einer Spezies die gleiche Farbe zugeordnet und der Baum als Abfolge von Farbsegmenten dargestellt. Für die TCC-Analyse ist es notwendig, wie für nahezu jede automatisierte Verarbeitung, einen Baum mit seiner Verzweigungsstruktur in einer linearen Form darzustellen. Die Notationsvereinbarung, die sich dabei durchgesetzt hat, drückt die Baumstruktur durch die Klammerung der beteiligten Elemente aus. Diese lineare Schreibweise eines Cluster-Baums als verschachtelte Listen bezeichnet man als *Newick-Repräsentation* des Baums. Der in Abbildung 1.10 dargestellte Baum dient der Erläuterung der Newick-Repräsentation. Die Knoten des Baums sind mit K_i bezeichnet, die von diesen Knoten ausgehenden Zweige besitzen die Längen L_i. Die Newick-Darstellung dieses Baums mit Zweiglängen lautet dann:

$$(K_4 : L_4, (K_1 : L_1, K_2 : L_2) : L_3) \; .$$

Dieser Baum kann nun in der Newick-Repräsentation durch die folgenden Ausdrücke dargestellt werden (hier ohne Zweiglängen):

$$(K_4, (K_2, K_1)) \quad \text{oder} \quad ((K_1, K_2), K_4) \quad \text{oder} \quad (K_4, (K_1, K_2)) \; .$$

An diesen drei unterschiedlichen (durch Permutation der geklammerten Elemente erzeugten) Darstellungen erkennt man, dass diese lineare Repräsentation eines Baums dieselben topologischen Freiheitsgrade besitzt wie der Baum selbst, d.h. die Reihenfolge der terminalen Knoten lässt sich im Einklang mit der Baumhierarchie (bzw. der Klammerung in der Newick-Darstellung) variieren. Die Sortierung des Baums basiert nun auf einem Algorithmus, bei dem im ersten Schritt die Information der Zweiglängen aus der Newick-Repräsentation gelöscht wird. Im zweiten Schritt wird, beginnend bei der äußersten Klammerung, überprüft, ob die betrachtete Liste als Element eine weitere Liste enthält. Diese Abfrage wird rekursiv fortgeführt, bis die betrachtete Liste ausschließlich einzelne Elemente und keine Liste mehr enthält. Der nächste Schritt besteht in der alphabetischen Sortierung dieser tiefsten Liste und ihrer Identifizierung durch das alphanumerisch erste Element. Die numerische Codierung der Nummer des Chromsoms wird dabei nicht berücksichtigt. Im letzten Schritt können von den Blättern beginnend in Richtung Wurzel nun sukzessiv alle höheren Listen alphabetisch sortiert werden. Der Algorithmus endet mit einem Baum, in dem die betrachteten Elemente so nah an einer universellen Reihenfolge sortiert sind, wie die Topologie (also die Verzweigungsarchitektur) des Baums es erlaubt. Diese Sortierung ändert nicht die Topologie des Baums. Abbildung 1.11 zeigt einen einfachen Baum und seine Sortierung mit Hilfe der *Tree Color Coding* Methode in drei verschiedenen Repräsentationen. Der Sortierungsalgorithmus neigt dazu, die Ordnung in einem Baum zu überschätzen, da Chromosomen einer Spezies auch dann direkte Nachbarn werden können, wenn diese in verschiedenen Zweiggruppen liegen und einer dieser Zweige Chromosomen einer anderen Spezies enthält. Ein solcher Effekt ist in dem Beispiel zum TCC-Algorithmus in Abbildung 1.11 zu sehen. Dort werden die Elemente die zu Spezies B gehören in der Farbabfolge direkt nebeneinander einsortiert, obwohl B_1 und B_2 zu verschiedenen Subclustern gehören. Wesentlich geringer ist dieser Effekt bei einer größeren Anzahl von Chromosomen einer Spezies.

1.5 $|t|$-Wert

Um zu untersuchen, wie die Information zur Speziestrennung innerhalb des Korrelationsvektors verteilt ist, wird ein Abstandsmaß definiert, das anders als eine Distanzmatrix den Beitrag einer Komponente des Vektors $\vec{\alpha}$ zur Trennung zweier Spezies quantifiziert. Komponenten von $\vec{\alpha}$, für die sich die Schar von Korrelationskurven zweier Spezies im Mittel stark unterscheiden, tragen stark zur Speziestrennung bei. Je größer die Varianz in einer Komponente α_k innerhalb einer Spezies ist, desto geringer trägt diese Komponente zur Trennung bei. In der Statistik wird der t-Wert verwendet, um eine Aussage über das Verhältnis von Unterschiedlichkeit zwischen zwei Gruppen, in diesem Fall zweier Spezies, treffen zu können. Der t-Wert ist folgendermaßen definiert:

$$t_k(A,B) = \frac{\bar{\alpha}_k(A) - \bar{\alpha}_k(B)}{\sqrt{\frac{\sigma_k^2(A)}{n(A)} + \frac{\sigma_k^2(B)}{n(B)}}}, \tag{1.19}$$

wobei $n(S)$ die Anzahl der Chromosomen der Spezies S beschreibt und $\bar{\alpha}_k(S)$ den Mittelwert über alle Korrelationskurven (also im Wesentlichen über alle Chromosomen) der Spezies S in der k-ten Komponente angibt. $\sigma_k^2(S)$ bezeichnet die Varianz in der Kurvenschar der Spezies S. Im Folgenden wird der absolute, auf die Summe Eins normierte t-Wert betrachtet und mit $|t|$ bezeichnet.

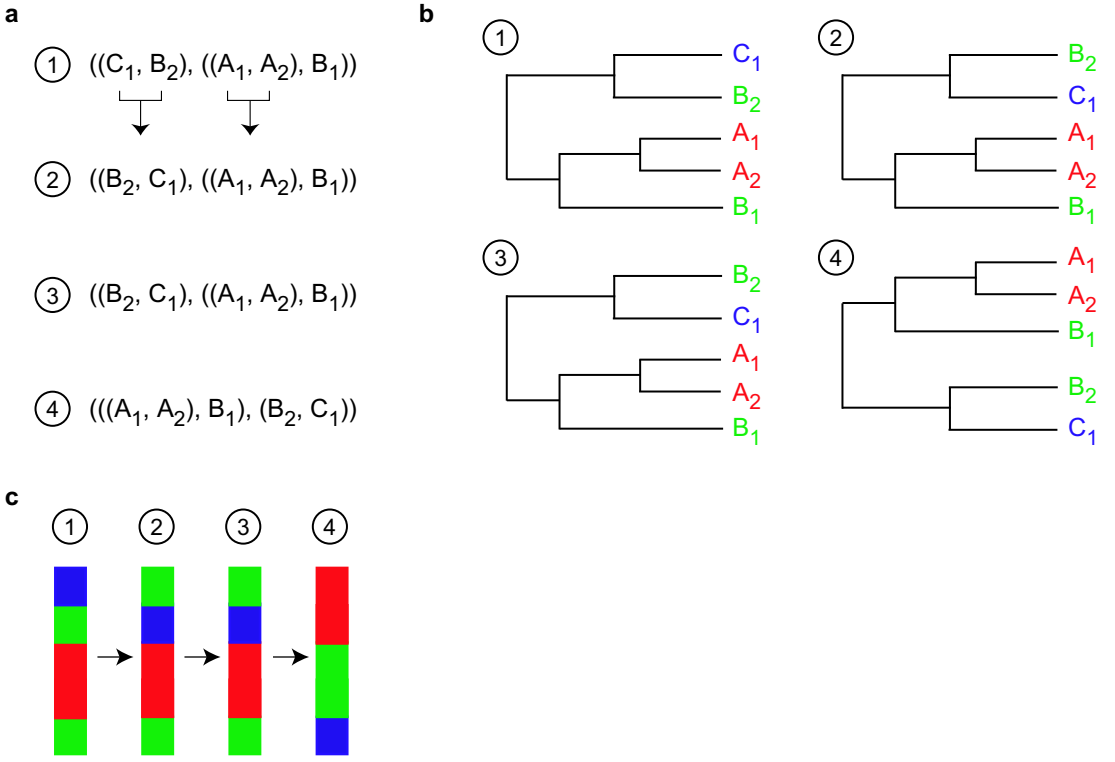

Abb. 1.11 a-c. Schematische Darstellung des *Tree Color Coding* (TCC) Algorithmus. **a** Vorgehensweise des TCC-Algorithmus auf Basis der Newick-Repräsentation für einen einfachen Baum mit fünf Taxa dreier verschiedener Spezies. Beginnend mit dem unsortierten Baum in (1) führt die Anwendung des TCC-Algorithmus durch iterative Vertauschung von Zweigen auf einen sortierten Baum (4). **b** Gleiche Operation wie in **a** auf Basis des Dendrogramms. **c** Visualisierung des Orginalbaums (1), der Zwischenschritte (2) und (3) und des finalen Baums (4) als TCC-Farbsegmente. (In Anlehnung an: Dehnert et al. (2006).)

1.6 Daten

Im Rahmen dieser Arbeit ist eine interne Genom-Datenbank angelegt worden, die regelmäßig auf einen aktuellen Stand gebracht wurde, indem aus den öffentlichen Genom-Datenbanken überarbeitete Versionen bereits nahezu vollständig sequenzierter Genome und neue, größtenteils sequenzierte Spezies eingepflegt wurden. Die drei wichtigsten primären Datenbanken *GenBank*[5], EMBL[6] und DDBJ[7] werden in kurzen zeitlichen Abständen synchronisiert und führen somit auf die gleiche Information bezüglich der Sequenzdaten. In Tabelle 1.1 findet sich eine Auflistung aller in dieser Arbeit untersuchten Spezies. Entscheidend für einen Vergleich der Daten sind die Versionsnummern der einzelnen Veröffentlichungen jedes Genoms. Im Anhang C findet sich eine Auflistung der in dieser Arbeit verwendeten Datensätze, spezifiziert durch die Angabe der Datenbank, der Versionsnummer und des Internetpfades zum Download der Daten. Außerdem ist

[5] *National Center for Biotechnology Information*, http://www.ncbi.nih.gov/Genbank/
[6] *European Molecular Biology Laboratory*, http://www.embl.org/
[7] *DNA DataBank of Japan*, http://www.ddbj.nig.ac.jp/

zu jeder Abbildung in dieser Arbeit dort eine Angabe der zugrunde liegenden Daten zu finden. (Tabelle C.4).

Die ersten publizierten Versionen des menschlichen Genoms im Februar 2001 durch das *Human Genome Sequencing Consortium* (2001) und ihrem kommerziellen Gegenstück, die durch Craig Venter geleitete Firma *Celera Genomics* (Venter et al., 2001), umfassen ca. 90% des euchromatischen Anteils des menschlichen Genoms, unterbrochen von ca. 150 000 nicht-annotierten Abschnitten (*gaps*) und haben eine Fehlerwahrscheinlichkeit von ca. 1 pro 10 000 Basen (Human Genome Sequencing Consortium, 2004). Als Gaps werden in diesem Zusammenhang also Abfolgen von Basen bezeichnet, die nicht identifiziert sind und durch ein einheitliches Symbol (N) dargestellt werden. Heterochromatin, das im Wesentlichen am Zentromer und an den Telomeren zu finden ist, lässt sich wegen seiner stark repetitiven Struktur nur sehr schwer sequenzieren. Auch im Euchromatin erschwert das Vorkommen von Sequenzwiederholungen und segmentellen Duplikationen die Bestimmung der Basenabfolge. Die im Jahre 2004 publizierte überarbeitete Version des menschlichen Genoms (HGSC Build 35) (Human Genome Sequencing Consortium, 2004) umfasst 99% des euchromatischen Anteils, weist 341 Gaps auf und besitzt eine Fehlerwahrscheinlichkeit von 1 pro 100 000 Basen. 33 Gaps (insgesamt ca. 198 Megabasen) gehen auf Heterochromatin zurück und 308 Gaps (insgesamt ca. 28 Megabasen) befinden sich im Euchromatin (Human Genome Sequencing Consortium, 2004). Damit ergibt sich rein rechnerisch eine durchschnittliche Länge von ca. 91 000 Basen pro Gap im Bereich des Euchromatin. Eine ähnlich hohe Datenqualität weisen z.B. die in dieser Arbeit untersuchten Genome von *Arabidopsis thaliana* (The Arabidopsis Genome Initiative, 2000), *Caenorhabditis elegans* (The C. elegans Sequencing Consortium, 1998) und *Drosophila melanogaster* (Celniker et al., 2002) auf. Allerdings wurden auch Genome die bisher nur als vorläufige Version (engl. *draft*) vorliegen, wie z.B. das Genom von *Gallus gallus* (International Chicken Genome Sequencing Consortium, 2004) analysiert. Der Einfluss der Datenqualität auf die in dieser Arbeit erfolgten Analysen und Ergebnisse, besonders die Auswirkungen der Anzahl von Gaps und deren prozentualer Anteil am Genom, wurden deshalb ausführlich untersucht. Dabei gibt es zwei Möglichkeiten, wie mit solchen nicht identifizierten Abschnitten verfahren werden kann. Die erste besteht in der Vernachlässigung solcher Bereiche, also dem Ausschneiden dieser Abschnitte, bestehend aus Ns, aus der Sequenz. Damit ergibt sich eine Verschiebung des Leserahmens. Eine alternative Vorgehensweise wird durch die Ersetzung der Gaps durch zufällige Sequenzen beschrieben, womit der Leserahmen erhalten bleibt. Bedingt durch die im Vergleich zum Gesamtvolumen relativ geringe Menge von nicht identifizierten Bereichen innerhalb eines Genoms, ist der Unterschied für beide Methoden in den Ergebnissen verschwindend klein. Wie im Falle des menschlichen Genoms beispielhaft beschrieben, gilt auch für die anderen in dieser Arbeit untersuchten Spezies, dass das Verhältnis von Anzahl der Gaps zur Gesamtmenge von nicht identifizierten Nukleotiden klein ist, es also relativ wenige, dafür aber große Gaps gibt. Dieser Sachverhalt legt eine Vernachlässigung dieser Bereiche nahe. Es lässt sich an dieser Stelle festhalten, dass die Ergebnisse äußerst robust gegenüber solchen Variationen in der Qualität der Daten sind, und dass keine signifikante Beeinflussung der Ergebnisse aufgrund unterschiedlicher Versionen von Datensätzen beobachtet wurde. Diese Validierungen wurden im Rahmen einer Diplomarbeit durchgeführt (Plaumann, 2003). In einem Fall wurde das Chromosom einer Spezies (*Gallus gallus*, Chromosom 16) aus der Analyse ausgeschlossen, da mehr als 20% des Chromosoms in der Veröffentlichung aus Gaps bestehen.

In dieser Arbeit wurde unter anderem untersucht, welche Auswirkungen einzelne Klassen von biologischen Komponenten (z.B. repetitive Elemente) auf die Korrelationsstruktur der DNA-

Tabelle 1.1. Auflistung aller in dieser Arbeit untersuchten Spezies und Informationen zu den Datensätzen in der lokalen Sequenzdatenbank. Zur weiteren Erläuterung sind hier allgemeine Bezeichnungen oder Trivialnamen der Spezies eingefügt. Auf diese Bezeichnungen wird im Laufe der Arbeit gelegentlich zurückgegriffen. Auf eine weitere systematische Charakterisierung oder eine präzise Artbezeichnung (z.B. Haushuhn, Wanderratte oder Malariamoskito) wurde hier verzichtet.

Spezies	**Allgemeine Bezeichnung**	**Anzahl veröff. Chromosomen**	**Sequenzlänge**
1. *Anopheles gambiae*	Moskito	4, X	228 Mbp
2. *Arabidopsis thaliana*	Acker-Schmalwand	5	119 Mbp
3. *Ashbya gossypii*	[Hefe]	7	9 Mbp
4. *Caenorhabditis elegans*	Fadenwurm	5, X	100 Mbp
5. *Cryptosporidium parvum*	[Parasit]	1	287 Kbp
6. *Danio rerio*	Zebrafisch	25	726 Mbp
7. *Drosophila melanogaster*	Taufliege	5, X	117 Mbp
8. *Encephalitozoon cuniculi*	[Parasit]	11	2 Mbp
9. *Gallus gallus*	Huhn	28	902 Mbp
10. *Homo sapiens*	Mensch	22, X,Y	3.070 Mbp
11. *Leishmania major*	[Parasit]	2	653 Kbp
12. *Mus musculus*	Maus	19, X	2.615 Mbp
13. *Oryza sativa*	Reis	2	46 Mbp
14. *Pan troglodytes*	Schimpanse	23, X,Y	3.084 Mbp
15. *Plasmodium falciparum*	[Parasit]	14	23 Mbp
16. *Rattus norvegicus*	Ratte	20, X	2.720 Mbp
17. *Saccharomyces cerevisiae*	Bäckerhefe	16	12 Mbp
18. *Schizosaccharomyces pombe*	Spalthefe	3	12 Mbp
19. *Trypanosoma brucei*	[Parasit]	1	1 Mbp

Sequenzen haben. Dafür ist es nötig, diese durch Positionsangaben in der Sequenz spezifizierten Bereiche aus der Analyse auszuschließen. Dies kann durch ein Überschreiben der Segmente durch zufällige Sequenzen erfolgen, was einem Löschen der Korrelationsstruktur entspricht. Die zweite Möglichkeit besteht in dem Ausschneiden diese Bereiche aus der Sequenz. In Abhängigkeit der Längenverteilung und der Menge an zu maskierender DNA werden im Folgenden beide Vorgehensweisen eingesetzt. Dabei wird deutlich, dass der Einfluss der Art der Maskierung sehr gering ist.

2

Ergebnisse und Diskussion

Das vorliegende Kapitel gliedert sich von den elementaren Korrelationsphänomenen aus zu den spezielleren und komplexeren Fällen. Das Prinzip der auf Korrelationen basierenden Genom-Signatur wird zuerst an drei klar trennbaren Spezies vorgeführt (Kapitel 2.1). Dann werden Spezies ergänzt, um die phylogenetische Dimension dieser Korrelationsanalysen ausleuchten zu können (Kapitel 2.2). Anhand einer Fallstudie zweier relativ eng verwandter Spezies mit hoher Verwandtschaft wird dann die Bedeutung des Symbolabstands und des Korrelationsbereichs diskutiert (Kapitel 2.2.3). Als nächstes wird die Analyse um zwei bezüglich ihrer Korrelationsstruktur schwer klassifizierbare Spezies erweitert (Kapitel 2.3). Diese Betrachtungen führen schließlich zu der Diskussion repetitiver DNA in ihrem Beitrag zu Korrelationen (Kapitel 2.4).

2.1 Speziesabhängigkeit der Korrelationskurven bei Mensch, Maus und Drosophila

Der Versuch, eine Spezies an Hand rein statistischer Eigenschaften ihrer DNA-Sequenz zu identifizieren, beschäftigt die Wissenschaft schon seit einigen Jahrzehnten. Die bekannteste solche Kenngröße einer Spezies bilden Korrelationen zwischen benachbarten Symbolen in Form von Dinukleotidhäufigkeiten (Russell et al., 1976; Russell und Subak-Sharpe, 1977; Karlin und Ladunga, 1994; Karlin und Mrázek, 1997; Karlin, 1998; Gentles und Karlin, 2001). Diese Größen spiegeln biologische Mechanismen und Prozesse wider, die auf der Ebene von Dinukleotiden wirken, wie zum Beispiel nachbarschaftsabhängige Mutationen (Arndt et al., 2002; Arndt und Hwa, 2005). Die Betrachtung von n-Wort Häufigkeiten ermöglicht eine Erweiterung des Spektrums in nur sehr geringem Maße (Hao und Qi, 2003; Qi et al., 2004), da für größere n (etwa ab $n > 5$) die Schätzung der Häufigkeiten für übliche Sequenzlängen zunehmend schwierig wird.[1] Die Korrelation zweier Symbole in einem Abstand k erlaubt, Mechanismen, Prozesse und Strukturen, die auf einer größeren Skala operieren oder vorhanden sind, statistisch zu erfassen und damit möglicherweise zu verstehen.

Die Informationstheorie stellt mit der Transinformation ein sehr gut geeignetes Werkzeug zur Verfügung, um solche Korrelationen zu quantifizieren. Dabei eignet sich dieses Maß für den Nachweis linearer und nichtlinearer Abhängigkeiten sowohl bei kleinen Symbolabständen als auch bei Symbolabständen über mehrere Größenordnungen (Herzel und Grosse, 1995, 1997; Grosse et al.,

[1] Bei größerem n müssten die Sequenzen unrealistisch lang sein, um den Möglichkeitsraum der n-Worte angemessen wiedergeben zu können.

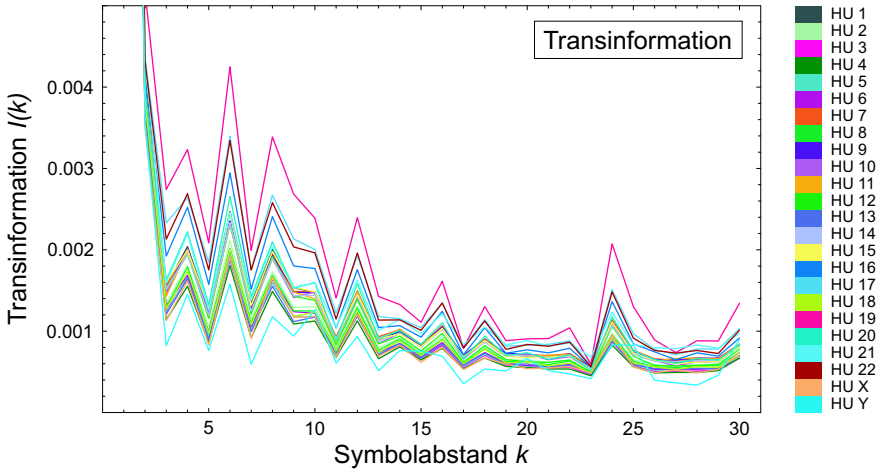

Abb. 2.1. Transinformation $I(k)$ im Symbolabstand k für die 22 Autosomen und die zwei Geschlechtschromosomen X und Y des *H. sapiens*. Die Zahl neben der Speziesabkürzung (HU) stellt die Nummer des jeweiligen Chromosoms dar.

2000; Holste et al., 2003). In Abbildung 2.1 ist die Transinformation $I(k)$ für die Chromosomen der Spezies *Homo sapiens* (Mensch) als Funktion des Abstandes k aufgetragen. Für jedes Chromosom wird dabei die Korrelationskurve für die volle Sequenzlänge bestimmt. Die Korrelationskurven der 22 Autosomen und der zwei Geschlechtschromosomen X und Y zeigen qualitativ einen ähnlichen Verlauf, sie sind jedoch nach oben verschoben. Das Chromosom 19 befindet sich am oberen Ende und das Chromosom Y am unteren Ende der Kurvenschar. Es ist deutlich zu sehen, dass die Abfolge der Maxima und Minima für alle Chromosomen gleich ist. Auch für das Y-Chromosom des Menschen trifft diese Aussage zu, wenn auch die Abweichungen von der Kurvenschar hierbei am deutlichsten sind. Dass die Korrelationskurven einen so ähnlichen Verlauf in der Transinformation zeigen, ist keinesfalls zu erwarten.

Die Quantifizierung der Korrelationsstärke durch die Parameter eines DAR(p)-Prozesses besteht aus einer Schätzung der Korrelationsstärke mit Hilfe einer empirischen Autokorrelationsfunktion, dem *ad hoc*-Schätzer, und der Bestimmung der Parameter durch die Lösung der Yule-Walker-Gleichungen (vgl. Kapitel 1.2.2 für eine ausführliche Darstellung). Abbildung 2.2 a zeigt die Kurven der empirischen Autokorrelationsfunktion für die Chromosomen des Menschen. Die Korrelationskurven der Chromosomen sind auch in diesem Maß sehr ähnlich, wobei die Verschiebung der Kurven zueinander in Richtung der Ordinate etwas geringer ausfällt als bei der Transinformation und somit eine leicht höhere Synchronisation eintritt. Das Y-Chromosom bildet hier eine Ausnahme. Während alle anderen Kurven die gleiche Abfolge von Maxima und Minima aufweisen, durchbricht Chromosom Y diese Abfolge für den Symbolabstand 15 und 20. Dies ist ein deutlicher Hinweis auf Unterschiede in den Korrelationsmaßen. Von der empirischen Autokorrelation des *ad hoc*-Schätzers gelangt man zu den Parametern des DAR(p)-Prozesses durch Lösung der Yule-Walker-Gleichungen. Diese Größen quantifizieren die Korrelationsstärke im Abstand k, wobei sie durch den DAR(p)-Prozess eine anschauliche Interpretation als Markov-Prozess $p$ter Ordnung besitzen (vgl. Kapitel 1.2.2) und werden hier auch als Markov-Repräsentation bezeichnet. Für die Chromosomen des Menschen sind die Korrelationskurven der Markov-Repräsentation

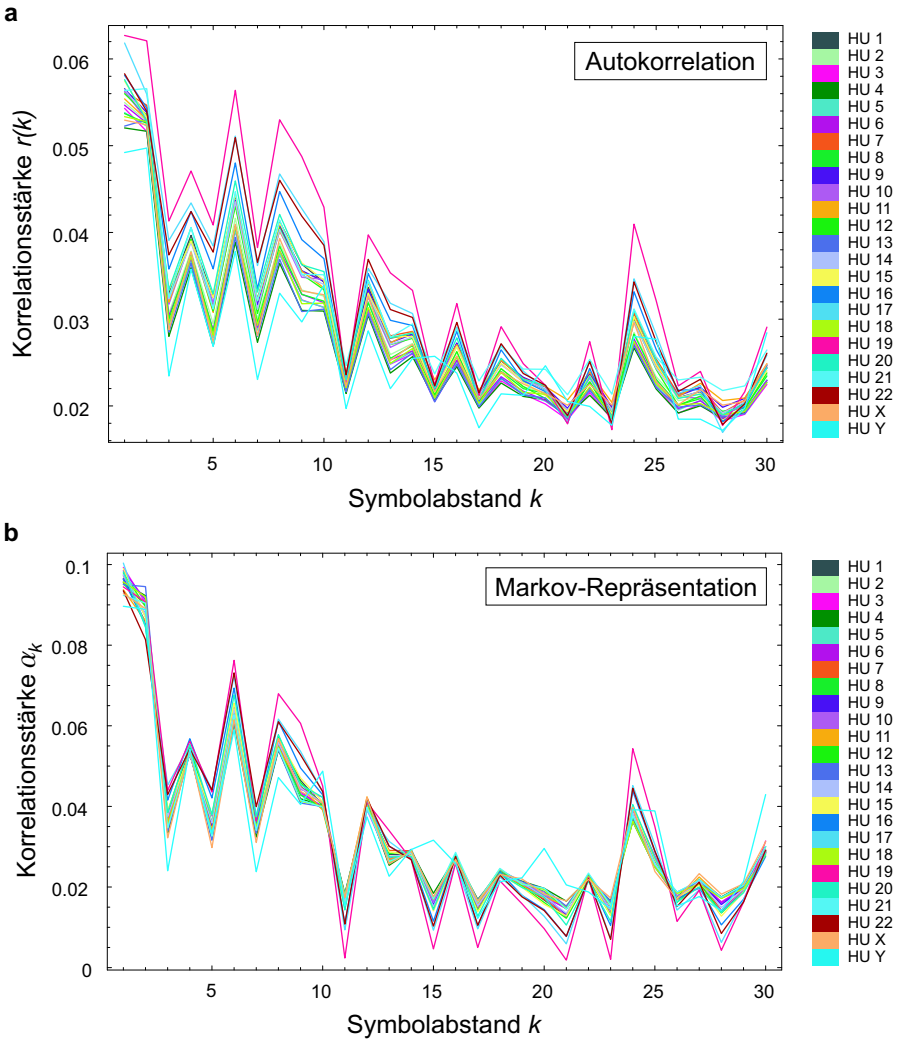

Abb. 2.2. **a** Empirische Autokorrelation $r(k)$ in Abhängigkeit des Symbolabstandes k für die 22 Autosomen und die zwei Geschlechtschromosomen X und Y des menschlichen Genoms. **b** Die aus den in Teil **a** angegebenen empirischen Korrelationskurven mit Hilfe der Yule-Walker-Gleichungen berechneten Parametervektoren $\vec{\alpha}$ (Markov-Repräsentation).

in Abbildung 2.2 b dargestellt. Die Korrelationskurven zeigen eine hohe Synchronisation. Sie liegen sehr nahe beieinander und weisen alle – bis auf das Y-Chromosom – die gleiche Struktur auf. Die Streuung in Richtung der Ordinate ist im Vergleich zu den Korrelationskurven des *ad hoc*-Schätzers deutlich reduziert. Eine solche Streuung kann mit dem unterschiedlichen globalen Gehalt an Korrelationen in Verbindung gebracht werden. Betrachten wir dazu noch einmal die Transinformationskurven aus Abbildung 2.1. Bedingt durch die unterschiedliche Menge an zufallsähnlichen Sequenzabschnitten innerhalb der verschiedenen Chromosomen streuen die Korrelationskurven stark in Richtung der Ordinate. Dieser Effekt ist auch bei den Korrelationskurven des *ad hoc*-Schätzers sichtbar, wenn auch weniger ausgeprägt. Die Reihenfolge der Kurven in Richtung der Ordinate bleibt dabei trotz der unterschiedlichen Maße im Vergleich zu Abbildung

2.1 und 2.2 a im Wesentlichen erhalten. Der Übergang zur Markov-Repräsentation führt zu einer Eliminierung dieser Streuung. Dies ist der zentrale Vorteil der Markov-Repräsentation gegenüber der Transinformation: Beim Schätzen der Parameter eines DAR(p)-Prozesses wird das zufällige „Rauschen“ in einer DNA-Sequenz, also der Hintergrund zufälliger Symbole, absorbiert und in Form des Parameters ρ quantifiziert.

Der nächste Schritt besteht in der Berechnung der Korrelationskurven für weitere Spezies. Abbildung 2.3 zeigt die Korrelationsstärke als Funktion des Symbolabstandes k für die Chromosomen der Spezies *Mus musculus* (Maus) und der Fruchtfliege *Drosophila melanogaster* (Drosophila). Für jedes Chromosom wird dabei die Korrelationskurve durch die Parameter eines DAR(30)-Prozesses quantifiziert, die aus den chromosomalen Sequenzen geschätzt werden. Die 19 Autosomen und das Geschlechtschromosom X der Maus zeigen eine ausgeprägte Oszillation der Periode zwei und eine sehr hohe Synchronisation. Die Korrelationskurven der fünf Autosomen und des X-Chromosoms von Drosophila besitzen eine andere Signatur als die der Maus, aber man sieht auch hier, dass die Korrelationskurven für alle Chromosomen von Drosophila einen sehr ähnlichen Verlauf aufweisen.

Nach der Betrachtung der Korrelationskurven von Mensch, Maus und Drosophila lässt sich festhalten, dass die Scharen von Korrelationskurven für diese Spezies starke qualitative Unterschiede aufweisen. Dieses Phänomen, die Synchronisation innerhalb einer Spezies und die unterschiedlichen Verläufe der Korrelationskurven für verschiedene Spezies, stellt das erste Ergebnis dieser Arbeit dar und bildet gleichzeitig den Ausgangspunkt für alle weiteren Untersuchungen.

2.1.1 Clusterbäume

Die hier angewandten Clustermethoden basieren auf paarweisen Abständen zwischen Korrelationskurven, die in Form von Distanzmatrizen zusammengefasst werden. Eine Beschreibung des genauen Verfahrens befindet sich in Kapitel 1.3. Das Bilden aller möglichen paarweisen Abstände führt zu einer symmetrischen Matrix, die in der Hauptdiagonalen Nullen aufweist. Die Einträge einer solchen Distanzmatrix können auf Werte zwischen Null und Eins normiert und in Form von Graustufen dargestellt werden. Dabei wird der Wert 0 durch die Farbe Weiß codiert und der Wert 1 durch die Farbe Schwarz. In Abbildung 2.4 ist die Distanzmatrix der Chromosomen des Menschen, der Maus und von Drosophila in Graustufen abgebildet. In dieser Darstellung zeigt sich die Ähnlichkeit der Chromosomen einer Spezies in vergleichbarer Weise wie in der Form der Korrelationskurven. Die Matrix ist von einem deutlichen Muster aus Flächen ähnlicher Graustufen geprägt. Die Intraspezies-Abstände von Chromosomen zeichnen sich durch helle Graustufen als Ausdruck eines geringen Abstands zwischen den Korrelationskurven aus. Dies ist bei der Maus am deutlichsten. Hier sieht man eine sehr homogene Struktur von hellen Graustufen innerhalb der Spezies. Bei Mensch und Drosophila ist diese Homogenität innerhalb der Spezies weniger ausgeprägt, was mit dem Eindruck, den man aus den Korrelationskurven gewonnen hat, in Einklang steht. Der Abstand zu dem X-Chromosom ist innerhalb von Drosophila am größten.

Die Interspezies-Abstände setzen sich durch dunklere Graustufen von den Intraspezies-Abständen ab und erlauben eine klare Trennung der einzelnen Spezies. Darüber hinaus sieht man, dass die größten Abstände sich dabei visuell für die Chromosomen von Drosophila zu denen des Menschen und der Maus in Form der dunkelsten Grauwerte ergeben. Die Abstände zwischen Mensch und

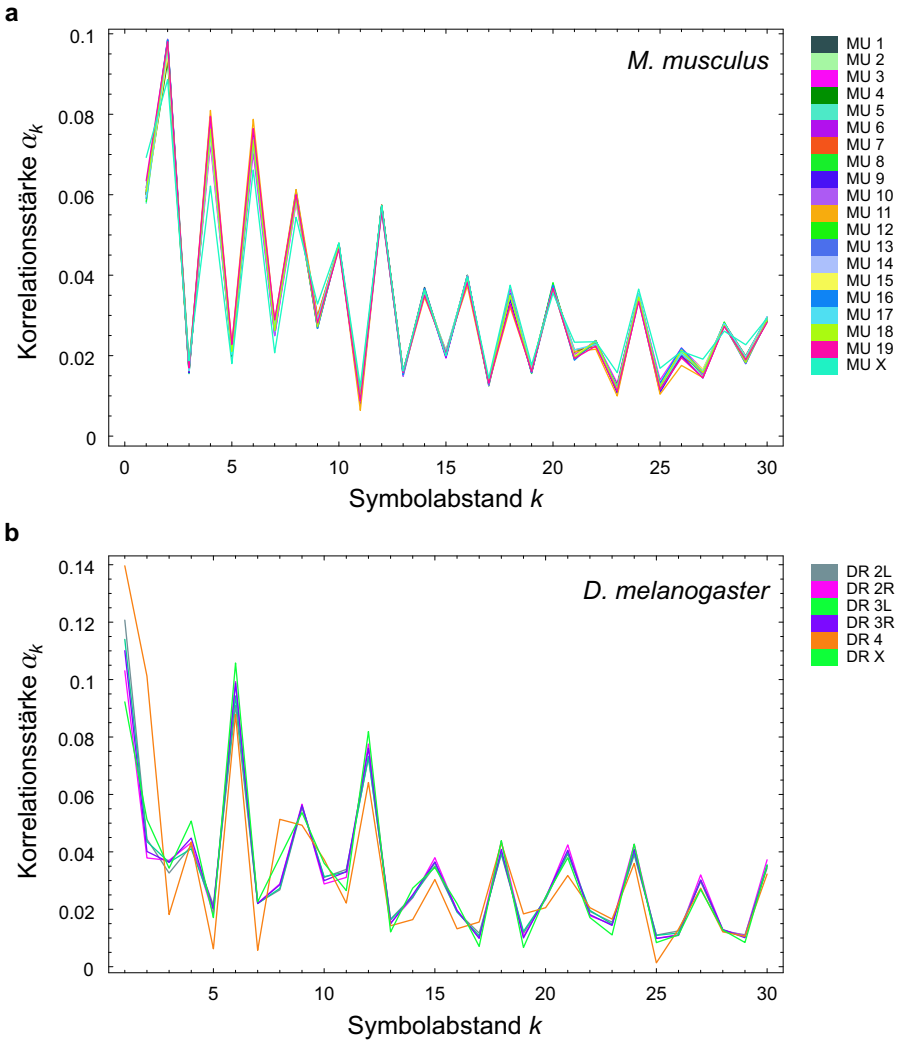

Abb. 2.3. Korrelationskurven für **a** die Chromosomen von *M. musculus* (MU) [20 Kurven] und **b** die Chromosomen von *D. melanogaster* (DR) [6 Kurven]. Die Korrelationskurven sind die aus den chromosomalen Sequenzen geschätzten Parametervektoren $\vec{\alpha}$ eines DAR(30)-Prozesses. (Angepasst aus: Dehnert et al. (2005b).)

Maus sind im Vergleich geringer. Diese Interpretation ist auf der Basis der Korrelationskurven ohne weitere Quantifizierung nicht möglich.

Die klare Trennung der Flächen und die Homogenität der Graustufen innerhalb der unterschiedlichen Rechtecke sind ein deutlicher Hinweis auf systematische Unterschiede in den Korrelationskurven der betrachteten Spezies. Der Clusteralgorithmus UPGMA kann nun die Distanzmatrix, die in Abbildung 2.4 in Graustufen codiert ist, in einen Baum übersetzen. Das Resultat ist in Abbildung 2.5 als Cluster-Baum dargestellt. Jedes Chromosom wird als eigenständiges Taxon betrachtet. Seine Spezieszugehörigkeit wird bei der Clusteranalyse nicht verwendet. Erst nach erfolgter Clusterung werden die Chromosomen farblich codiert, um eine transparente Darstellung zu unterstützen. Die Clusteranalyse ist eine sehr effiziente Methode, die in den Korrelationskurven

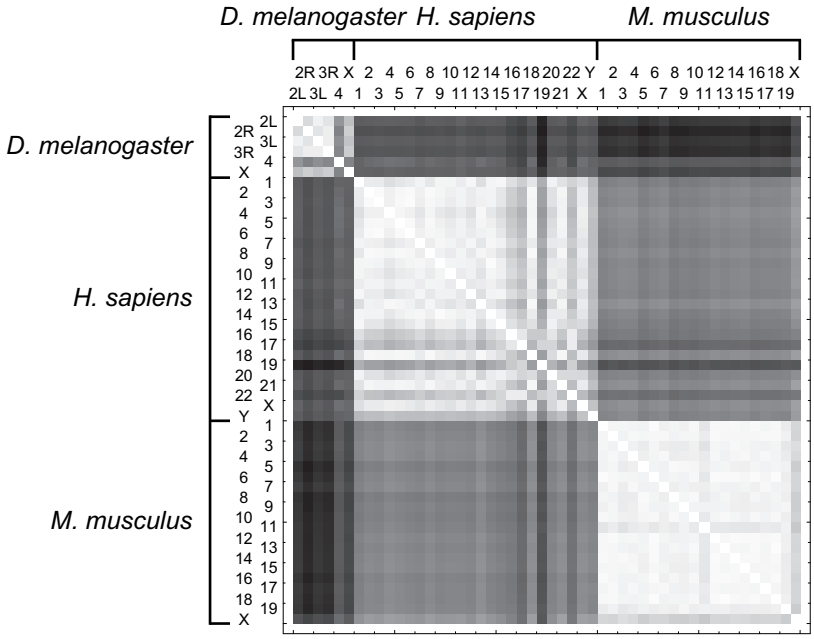

Abb. 2.4. Darstellung der Distanzmatrix für die Chromosomen von *D. melanogaster*, *H. sapiens* und *M. musculus* in Graustufen.

enthaltene Information in eine relationale Struktur zu übersetzen. Die Chromosomen der einzelnen Spezies entsprechen in diesem Baum den Endknoten oder den Blättern, die sich am Ende der Zweige befinden und in der Clusteranalyse als eigenständige Elemente betrachtet werden. Die Länge der Zweige oder Kanten des Baums codiert den Abstand der Knoten. Die Chromosomen bilden dabei Subcluster die sich auf der Ebene der Spezies klar abgrenzen. Es entstehen so drei Cluster die ausschließlich Chromosomen einer Spezies beinhalten. Betrachtet man nun die interne Struktur der Speziescluster, so zeigen die Chromosomen der Maus die geringsten Zweiglängen innerhalb des Clusters. Der größte Abstand besteht zwischen den Autosomen und dem Geschlechtschromosom X, das als erstes abzweigt. Die Kantenlängen innerhalb des Clusters der Chromosomen des Menschen sind größer und es bilden sich deutliche Subcluster von Chromosomen. Bei Drosophila liegt Chromosom 4 am Rand des Clusters, noch vor dem X-Chromosom, das sich wiederum von den vier verbleibenden Autosomen absetzt. Diese Darstellung in Form eines Baums zeigt erstmals eine definitive Clusterung der Chromosomen einer jeden Spezies auf Basis der Korrelationskurven.

2.2 Erweiterung der Analyse um *C. elegans*, Moskito und Ratte

Der nächste Schritt liegt in der Erweiterung des Spektrums von Spezies. Dafür werden die Korrelationskurven von *Anopheles gambiae* (Moskito), *Caenorhabditis elegans* (Fadenwurm) und *Rattus norvegicus* (Ratte) aus den chromosomalen Sequenzen bestimmt und in die Clusteranalyse einbezogen. In Anhang B in Abbildung B.1 sind die Korrelationskurven für die nun sechs betrachteten Spezies dargestellt. Der in Abbildung 2.6 dargestellte Baum basiert erneut auf der L_1-Norm in

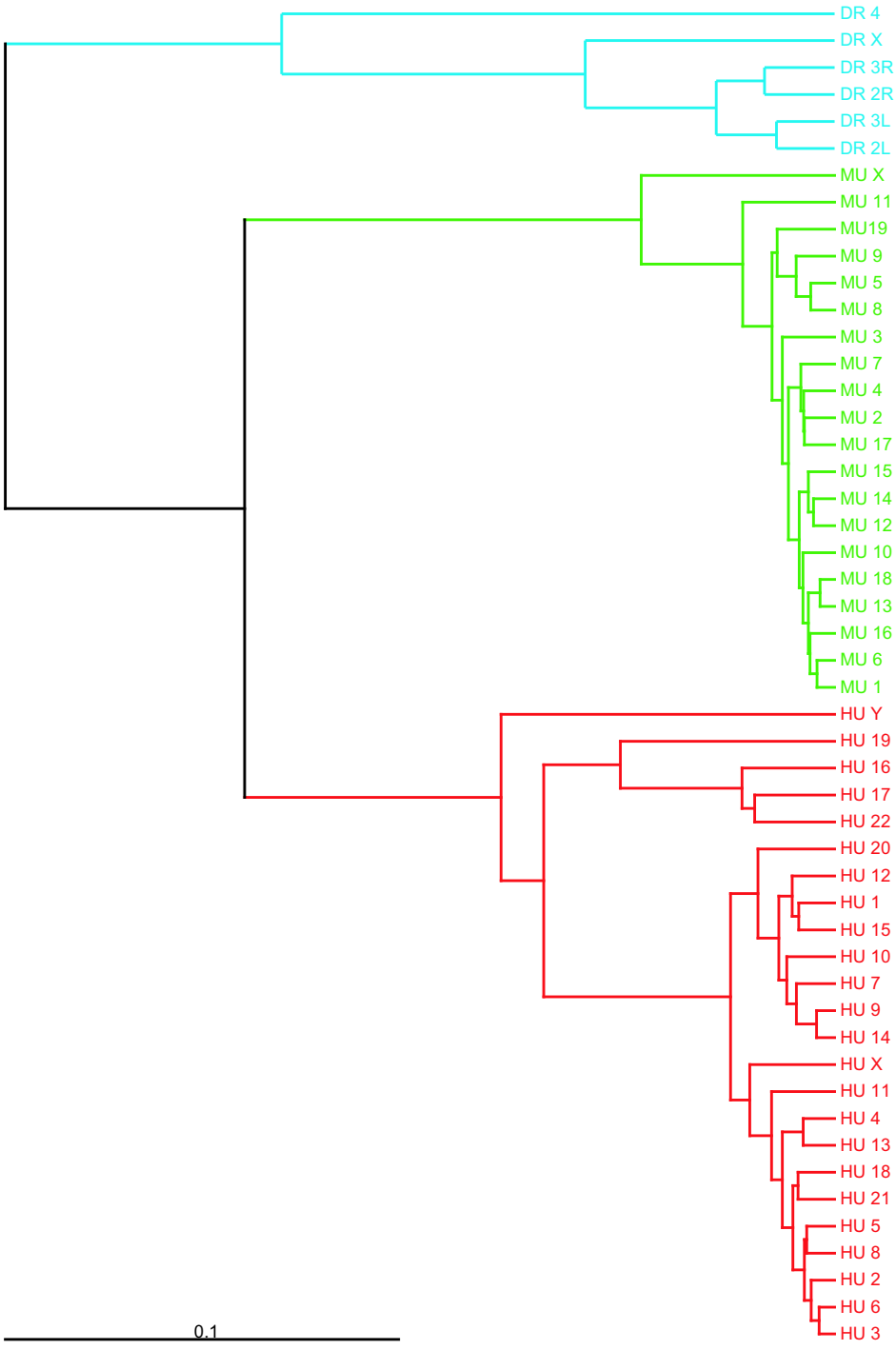

Abb. 2.5. Clusterbaum für die Chromosomen von *D. melanogaster* (DR), *H. sapiens* (HU) und *M. musculus* (MU) durch Anwendung des UPGMA-Algorithmus auf die in Abbildung 2.4 in Graustufen dargestellte Distanzmatrix. Die Zahl neben der Speziesabkürzung stellt die Nummer des jeweiligen Chromosoms dar. Die Legende ermöglicht die Vergleichbarkeit der Zweiglängen.

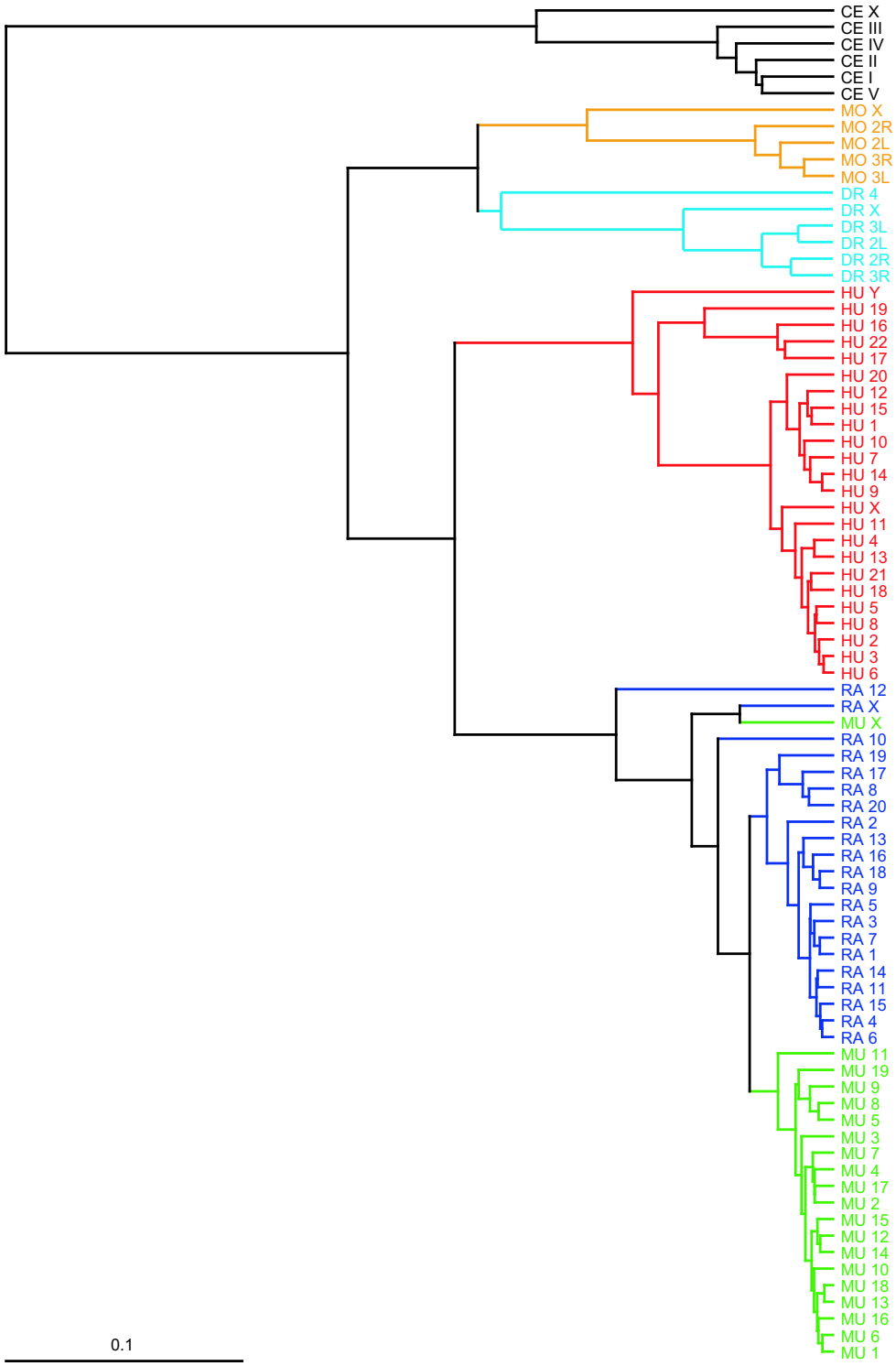

Abb. 2.6. Clusterbaum für sechs eukaryotische Spezies basierend auf der Markov-Repräsentation. Die betrachteten Spezies sind: *A. gambiae* (MO); *C. elegans* (CE); *D. melanogaster* (DR); *H. sapiens* (HU); *M. musculus* (MU); *R. norvegicus* (RA).

Verbindung mit dem Clusteralgorithmus UPGMA. Die Chromosomen der Spezies bilden Cluster, und es ergibt sich eine klare Trennung der Chromosomen von *C. elegans*, Drosophila, Moskito und Mensch. Die Cluster der Chromosomen von Ratte und Maus fallen eng zusammen, sie bilden jedoch zugleich große reine Subcluster aus Chromosomen der jeweiligen Spezies. Die Chromosomen an der Wurzel dieser Subcluster sind die Geschlechtschromosomen X der Ratte und Maus, sowie das Chromosom 12 der Ratte. Die Positionierung von Geschlechtschromosomen am äußeren Rand der jeweiligen Chromosomencluster ist systematisch. So liegt für *C. elegans*, Moskito und Drosophila das X-Chromosom am jeweiligen Rand der Cluster und für den Menschen das Y-Chromosom.

Neben der Clusterung der Chromosomen einer Spezies sieht man sofort, dass die Struktur des Baums auch phylogenetische Aspekte widerspiegelt. Die nahe Verwandtschaft von Maus und Ratte, sowie von Drosophila und Moskito findet sich ebenso wieder wie die Unterscheidung von Säugetieren und Insekten in Abgrenzung zu *C. elegans*. Diese Lesart des Baums ist nicht phylogenetisch, sondern beschreibt Relationen und Verwandtschaften. Besonders deutlich wird dies bei der Betrachtung der Cluster von Chromosomen. Eine phylogenetische Interpretation im Sinne gemeinsamer Vorfahren ist dort nicht sinnvoll. Auf der Ebene klassischer Clusteranalysen lässt sich diese Struktur aber sehr gut interpretieren. So zeigen z.B. die Chromosomen 10, 7, 14 und 9 des Menschen Gemeinsamkeiten, wobei diese zwischen Chromosom 14 und 9 am größten sind (siehe Abbildung 2.6).

Auch ohne weitere Analysen lässt sich vermuten, dass diese qualitativen Übereinstimmungen mit einer phylogenetischen Systematik ihren Ursprung in Prozessen der Genom-Evolution haben: Unterschiede auf dieser statistischen Ebene (also in den Korrelationskurven) können sich umso stärker aufbauen, je länger der Zeitpunkt der entsprechenden Speziesdifferenzierung zurückliegt.

Die gleiche Analyse kann nun auch mit der Transinformation als Korrelationsmaß durchgeführt werden. Abbildung 2.7 zeigt den resultierenden Baum. Auch hier sieht man wie in Abbildung 2.6 eine deutliche Clusterung der Chromosomen einer jeden Spezies. Die Chromosomen von C. elegans werden jedoch als einzige vollständig zusammengefasst, ohne einen Einschluss von Chromosomen einer anderen Spezies. In allen anderen Fällen bilden sich große Subcluster von Chromosomen, die ein oder mehrere Chromosomen einer anderen Spezies enthalten oder diese gemeinsam teilen. Die X-Chromosomen von Drosophila und Moskito, sowie das Chromosom IV von Drosophila zweigen von den Clustern der restlichen Chromosomen dieser beiden Spezies ab. Noch größer ist der Unterschied bei den verbleibenden drei Spezies im Vergleich zu Abbildung 2.6. Die Chromosomen von Mensch und denen der Maus und der Ratte werden nicht vollständig separaten Clustern zugewiesen. Die Chromosomen 19, 16, 22 und 17 des Menschen liegen vor den Clustern dieser Spezies. Die restlichen Chromosomen des Menschen bilden ein homogenes Cluster. Während bei der Markov-Repräsentation in Abbildung 2.6 die Chromosomen der Maus und Ratte fast vollständig getrennt werden, ist dies bei der Transinformation nicht der Fall. Die Chromosomen bilden Subcluster, von denen eine größere Anzahl von Chromosomen beider Spezies abzweigt, darunter auch die Geschlechtschromosomen X der Maus und Ratte. Aber auch hier bleibt eine Trennung von Säugetieren, Insekten und *C. elegans* erhalten. Der Vorteil der Markov-Repräsentation gegenüber der Transinformation in Bezug auf die Clusterung der Chromosomen ist nach diesem Vergleich offensichtlich.

Die gleiche Untersuchung wurde auch mit unterschiedlichen Kombinationen von Distanzmaß und Clusteralgorithmus durchgeführt. Dabei führte die L_1-Norm in Verbindung mit UPGMA auf ver-

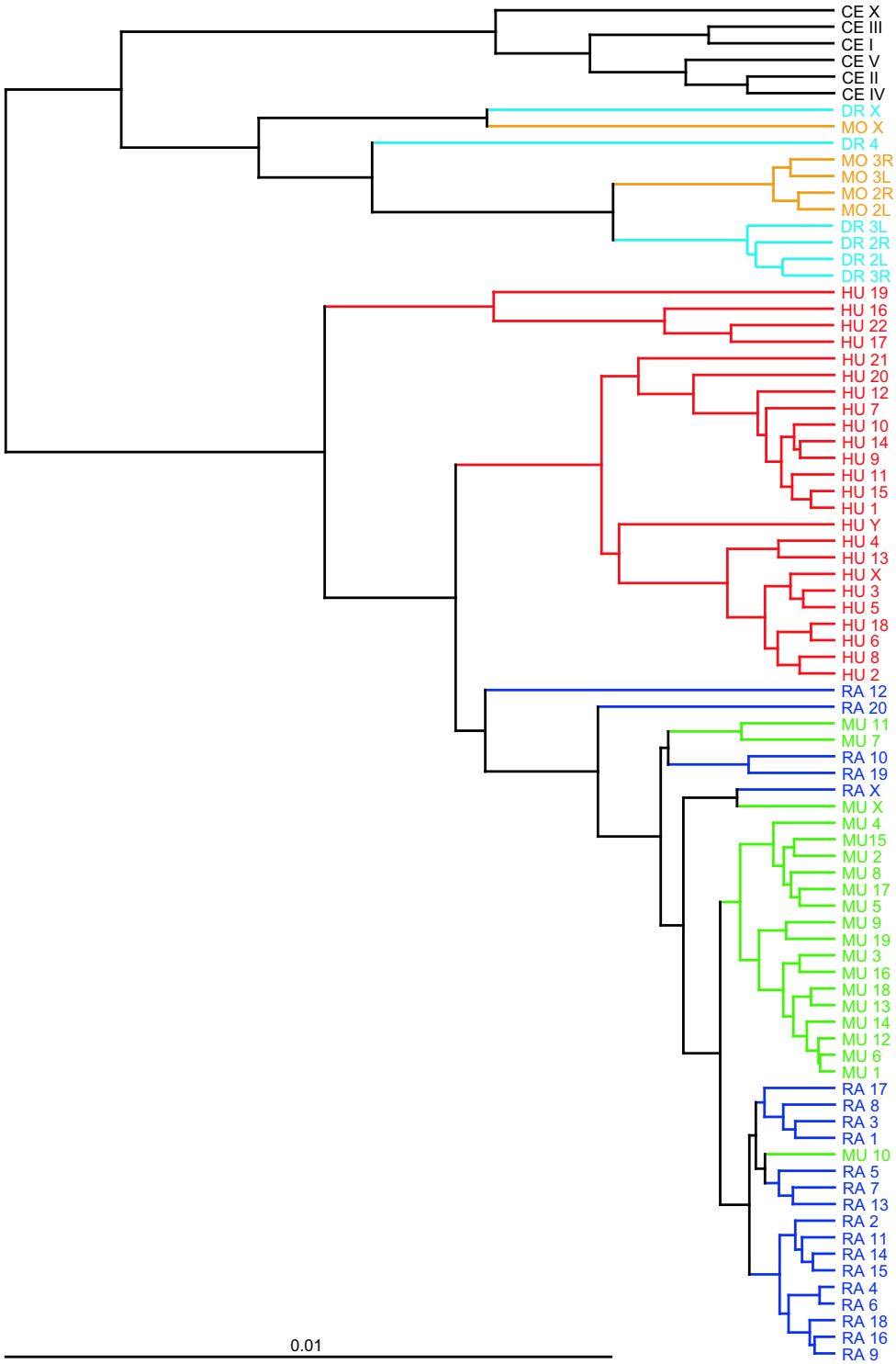

Abb. 2.7. Clusterbaum für sechs eukaryotische Spezies basierend auf der Transinformation als Korrelationsmaß. Die betrachteten Spezies sind: *A. gambiae* (MO); *C. elegans* (CE); *D. melanogaster* (DR); *H. sapiens* (HU); *M. musculus* (MU); *R. norvegicus* (RA).

gleichsweise gute Ergebnisse. Ein weiterer in der Phylogenie oft eingesetzter Clusteralgorithmus ist das Neighbour-Joining Verfahren, das für die Markov-Repräsentation auf einen qualitativ ähnlichen Baum führt wie das UPGMA-Verfahren. Dieser Aspekt wurde in einer Diplomarbeit untersucht (Krauss, 2006).

2.2.1 Robustheit der Bäume

Die Robustheit der in Abbildung 2.6 und 2.7 dargestellten Bäume gegenüber einer Variation der zugrunde liegenden Daten oder einer Änderung der Parameter der Analyse stellt einen wichtigen methodischen Untersuchungsgegenstand dar. Die Stabilität eines Clusterbaums kann mit Bootstrap-Verfahren überprüft werden, indem ein Teil der vorhandenen Information systematisch bei der Konstruktion des Baums weggelassen wird (siehe Kapitel 1.3.3). Wird dieses Verfahren unter Vernachlässigung verschiedener Teilinformationen wiederholt, ergibt sich ein Bild davon, wie robust die einzelnen Verzweigungen gegenüber solchen Manipulationen sind. Das Weglassen einzelner Komponenten zweier Korrelationskurven bei der Berechnung des paarweisen Abstandes stellt ein solches Vorgehen dar. Der in Abbildung 2.8 dargestellte Baum ist die Übereinstimmung von 100 Bäumen, bei deren Konstruktion zufällig sechs bzw. 20% der Komponenten der Korrelationsvektoren bei der Bestimmung der Distanz zueinander vernachlässigt wurden. Die Zahlen an den Verzweigungen stellen die Bootstrap-Werte dar, die angeben wie häufig dieser Knoten in 100 Bootstrap-Bäumen vorhanden ist. Die Knoten zwischen den Clustern der einzelnen Spezies weisen sehr hohe Bootstrap-Werte auf, was auf einen robusten Baum hindeutet. Die Verzweigung zwischen den Chromosomen des Menschen und denen der Maus und Ratte hat den Bootstrap-Wert 100 und ist somit in allen 100 Bäumen an dieser Stelle. Auch die Verzweigung zwischen den Insekten und den Säugetieren hat mit 97 einen sehr hohen Bootstrap-Wert. Selbst die Aufteilung zwischen den Chromosomen der Maus und Ratte in die zwei großen Subcluster mit einem Bootstrap-Wert von 73 ist noch relativ stabil gegenüber einer solchen Manipulation der Daten. Diese Analyse zeigt, dass das zufällige Weglassen einzelner Komponenten qualitativ auf den gleichen Baum führt und somit keine stochastischen Effekte für die Clusterung verantwortlich sind, sondern Informationen die über die gesamte Korrelationskurve verteilt sind. Es ist zu beachten, dass die Bäume in Abbildung 2.6 und 2.8 mit unterschiedlichen Parametereinstellungen der Clusteralgorithmen bestimmt worden sind. Der mit Bootstrap-Werten versehene *Consensus*-Baum (Abbildung 2.8) kann so leichte Unterschiede zu dem Einzelbaum (Abbildung 2.6) aufweisen; vgl. auch Kapitel 1.3.3 Abbildung 1.9.

2.2.2 Längenabhängigkeit

Die hier vorgestellte Methode der Konstruktion von Clusterbäumen auf Basis der Korrelationskurven hat mehrere freie Parameter, die sich auf die Ergebnisse der Analyse auswirken. Die Sequenzlänge der zugrunde liegenden Daten stellt einen solchen, sehr wichtigen Parameter der Analyse dar. Für die vollständigen chromosomalen Sequenzen erhält man die in den bisher betrachteten Abbildungen visualisierten Bäume. Um die Clusterung der Chromosomen in Abhängigkeit der Sequenzlänge zu untersuchen, ist es nötig, eine geeignete Visualisierung zu finden, die es erlaubt eine große Anzahl von Bäumen kompakt und vergleichend darzustellen. Das im Rahmen dieser Arbeit entworfene Verfahren des *Tree Color Coding* stellt dabei einen Clusterbaum als Farbabfolge

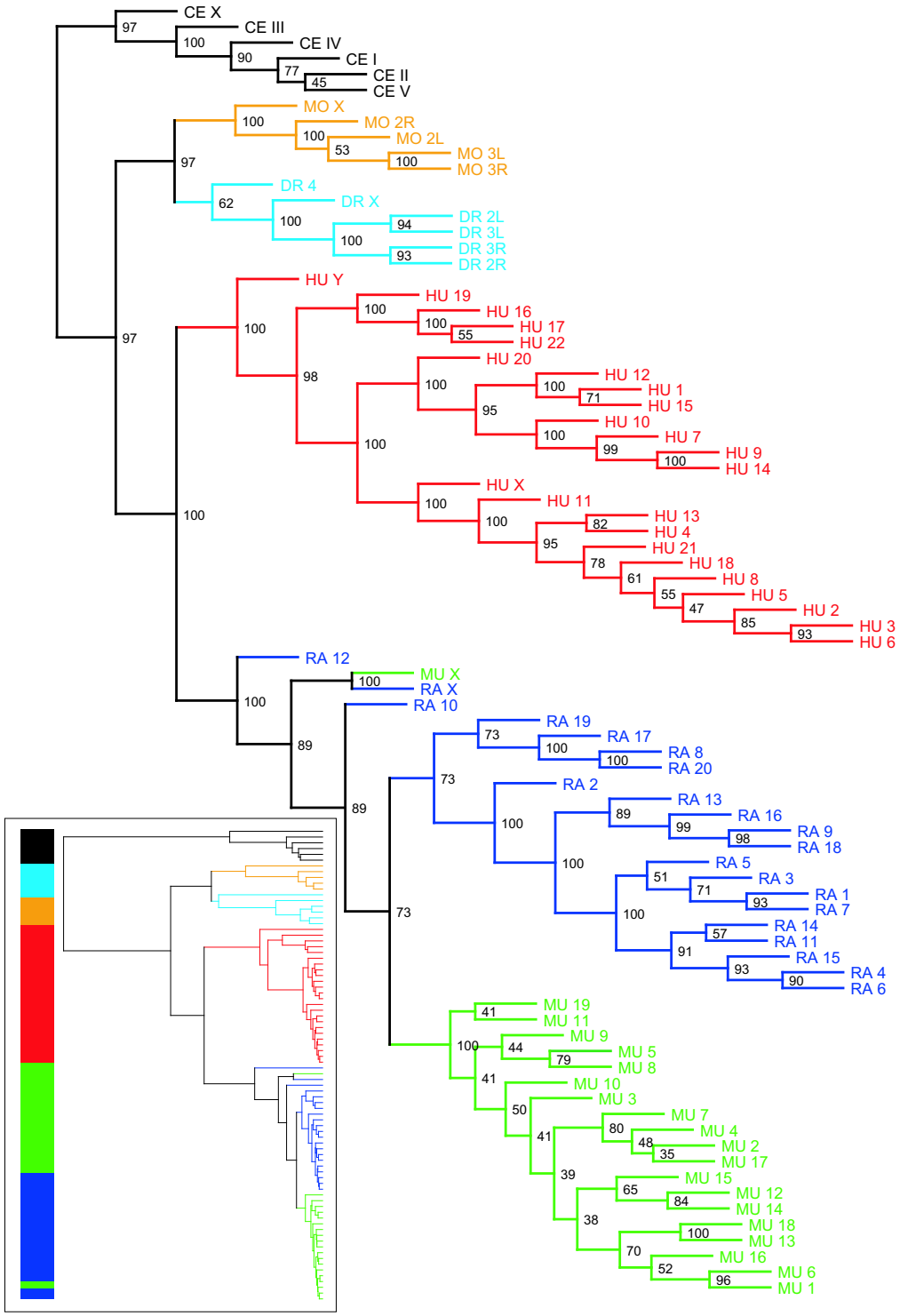

Abb. 2.8. Clusterbaum für sechs eukaryotische Spezies. Der Baum stellt den *Consensus*-Baum für 100 Bootstrap-Samples dar. Die Zahlen an den Knoten entsprechen den Bootstrap-Werten. Die betrachteten Spezies sind: *A. gambiae* (MO); *C. elegans* (CE); *D. melanogaster* (DR); *H. sapiens* (HU); *M. musculus* (MU); *R. norvegicus* (RA). (Aus: Dehnert et al. (2005b).)

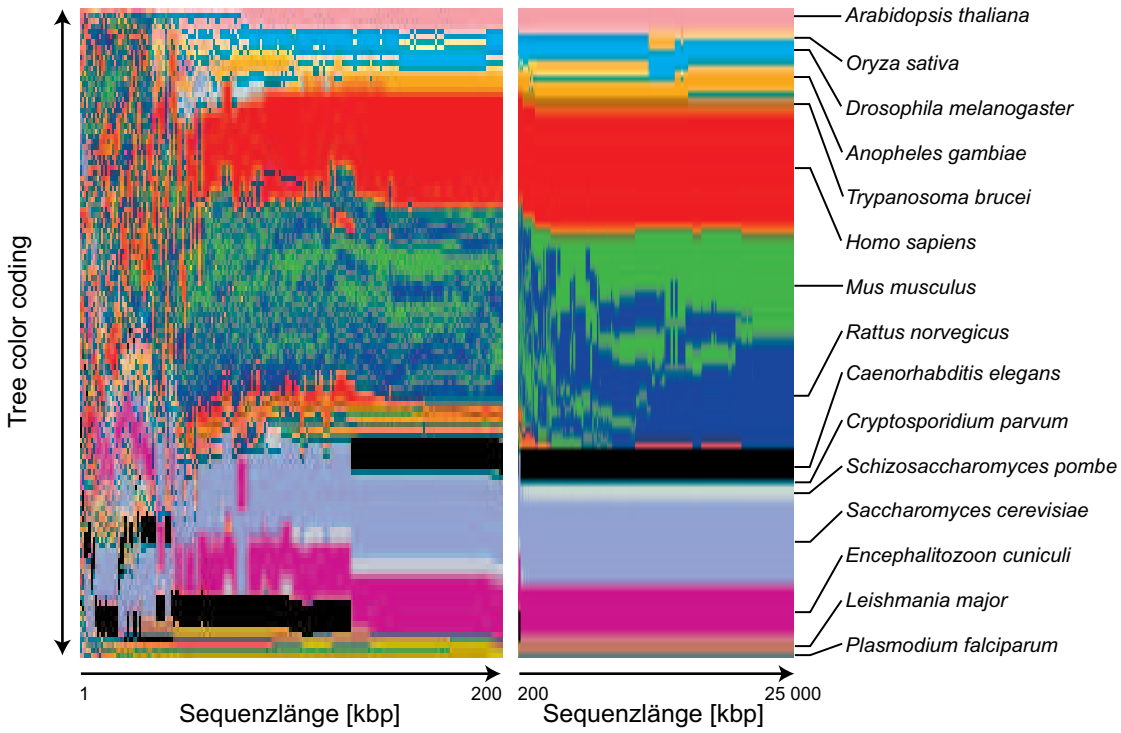

Abb. 2.9. *Tree Color Coding* Plot für 15 eukaryotische Spezies. Der Einfluss der Länge der zugrunde liegenden Sequenzen auf die Clusterung der Chromosomen wird visualisiert. (Aus: Dehnert et al. (2005b).)

dar. Auf diese Weise lassen sich die Auswirkungen einzelner Parameter der Analyse quasikontinuierlich in der Clusterung des Baums beobachten. Das genaue Vorgehen des TCC-Algorithmus ist in Kapitel 1.4 beschrieben. Abbildung 2.8 zeigt im Einsatz (links unten) den einfachen Clusterbaum ohne Bootstrap-Analyse zusammen mit der Codierung als Farbabfolge.

Abbildung 2.9 zeigt die Längenabhängigkeit einer Clusteranalyse mit 15 Spezies und 124 Chromosomen in Form eines TCC-Plots. Die Korrelationskurven wurden auf der Basis einer sukzessiv um 1000 Basen (bp)[2] vergrößerten Teilsequenz eines jeden Chromosoms berechnet. Beginnend mit den ersten 1000 Basen einer jeden Sequenz und einer Schrittweite von 1000 Basen wird dies bis zum Erreichen von 25 000 kbp fortgesetzt und dabei der jeweilig erhaltene Clusterbaum in eine Farbabfolge übersetzt. Chromosomen, die ihre volle Länge erreicht haben, werden ab diesem Zeitpunkt in ihrer Gesamtlänge berücksichtigt.

Für sehr kurze Sequenzen ergibt sich eine durchmischte Farbabfolge. Die durch die Korrelationskurven aus den Sequenzen extrahierte Information ist nicht ausreichend, um systematische Gruppen von Chromosomen zu bilden. Mit größer werdender Sequenzlänge bilden sich die ersten größeren Cluster von Chromosomen und man erkennt, dass schon bei 100 kbp eine gewisse Ordnung im Baum existiert. Für die klare Trennung der in dieser Analyse am engsten verwandten

[2] bp = Basen*paare*. Diese übliche Einheit, die sich auf doppelsträngige DNA bezieht, wird hier verwendet, obwohl in die vorliegenden Analysen Einzelstränge einfließen.

Spezies, nämlich Maus und Ratte, wird die größte Sequenzlänge benötigt. Die evolutionäre Distanz der beiden Spezies reicht offensichtlich gerade aus, um bei voller Sequenzlänge aufgrund des Korrelationsprofils zu einer klaren Unterscheidung der Spezies zu gelangen. Als wichtiges Ergebnis lässt sich damit festhalten, dass die Länge der DNA-Sequenzen eine entscheidende Rolle bei der Analyse darstellt, und dass der *Tree Color Coding* Plot eine transparente Darstellung des Einflusses dieses Parameters auf die Ergebnisse erlaubt.

2.2.3 Fallstudie: Maus und Ratte

Die Eigenschaft von Maus und Ratte, erst bei einer im Verhältnis großen Sequenzlänge getrennte Cluster von Chromosomen zu bilden, machen sie zu idealen Untersuchungsobjekten in Bezug auf die Parameter der Analyse. Von besonderem Interesse ist dabei der betrachtete Bereich von Symbolabständen bei der Berechnung der Korrelationsstärke. Betrachten wir aus diesem Grund diese beiden Spezies etwas genauer. Abbildung 2.10 zeigt die Korrelationskurven bis zu einem Symbolabstand von $p = 50$ für die Maus (a) und Ratte (b), wobei die Geschlechtschromosomen nicht dargestellt sind. Die bisherigen Analysen in diesem Kapitel haben gezeigt, dass die Geschlechtschromosomen der hier untersuchten Spezies oft eine deutlich abweichende Korrelationsstruktur aufweisen und sich folglich am Rand der Speziescluster wiederfinden. Da in der vorliegenden Arbeit die Analyse speziestypischer statistischer Eigenschaften im Vordergrund steht, werden die Geschlechtschromosomen in allen folgenden Untersuchungen dieser Arbeit nicht berücksichtigt. Die Korrelationskurven der Maus in Abbildung 2.10 a zeigen für $k = 1, \ldots, 50$ ein hoch synchronisiertes Verhalten und sind von denen der Ratte in Abbildung 2.10 b visuell schwer zu unterscheiden. Auf dieser Basis ergibt sich kein Unterschied zwischen den Kurven der Maus für $p = 50$ und denen für $p = 30$ aus Abbildung 2.2 b. Ob ein systematischer Unterschied zwischen den Kurvenscharen von Maus und Ratte existiert, lässt sich qualitativ in Form einer Stichprobe durch das Auftragen weniger Kurven in einer gemeinsamen Graphik überprüfen. In Abbildung 2.10 c sind jeweils die Korrelationskurven der Chromosomen 1 und 2 aufgetragen und man erkennt Unterschiede, die für einzelne Abstände k in ihrer Größe variieren. Als Ausschnitt ist der Bereich zwischen 24 und 36 dargestellt, in dem systematische Abweichungen für $k = 26$ und $k = 29$ sichtbar sind.

Die Berechnung der Distanzmatrix basierend auf der L_1-Norm und ihre Darstellung in Abbildung 2.11 als Graustufenwerte verdeutlichen die Unterschiede bei Betrachtung aller Chromosomen beider Spezies. Es bilden sich zwei Klassen von Graustufen, helle im Bereich der Intraspezies-Abstände und dunkle für die Interspezies-Abstände. Dabei erkennt man jedoch auch einzelne dunkle Bereiche bei den Intraspezies-Abständen. Dies ist für die Chromosomen 10 und 12 der Ratte und auch für Chromosom 11 der Maus der Fall. Hier zeigen sich dunkle Graustufen in Form von Balken die sich durch die Matrix ziehen und einen großen Abstand visualisieren.

Es ist nach diesen Betrachtungen zu erwarten, dass für $p = 50$ eine ähnliche Trennung der Chromosomen beobachtet werden kann wie für $p = 30$. Abbildung 2.12 zeigt den unter der Verwendung des UPGMA-Algorithmus gewonnenen Clusterbaum mit Bootstrap-Werten. Die Chromosomen 10 und 12 der Ratte sowie das Chromosom 11 der Maus liegen − wie erwartet − vor den eigentlichen Clustern der Chromosomen, und es ergibt sich ein ähnliches Bild wie für $p = 30$ in Abbildung 2.6 für das Subcluster von Maus und Ratte. Die Erweiterung des Abstandsbereiches k bei der Berücksichtigung der vollen Sequenzlänge führt auf dieser Ebene zu keiner signifikanten Verbesserung der Clusterung der Chromosomen.

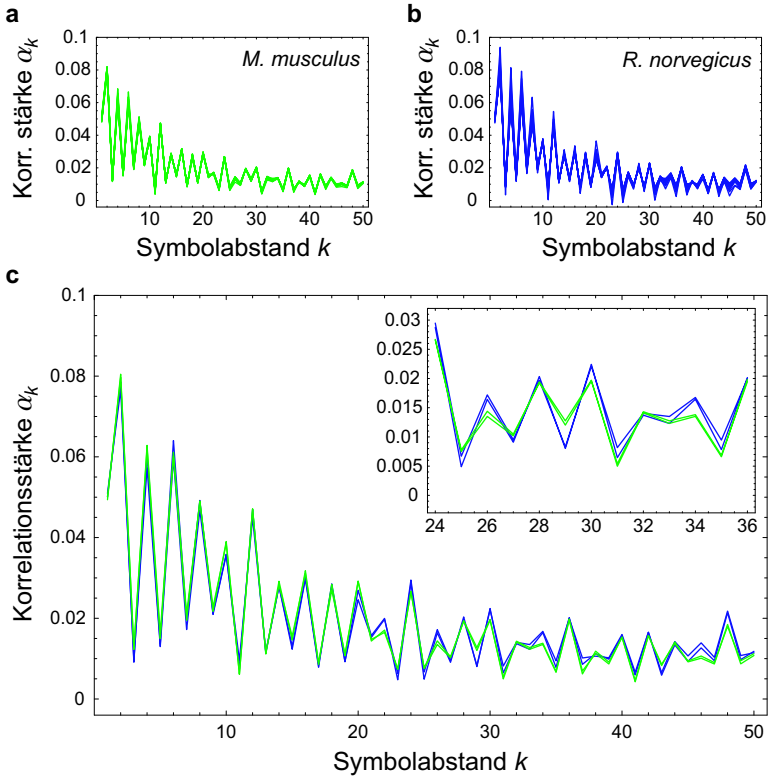

Abb. 2.10. Korrelationskurven der Markov-Repräsentation der Ordnung $p = 50$ für **a** die Autosomen von *M. musculus* und **b** die Autosomen von *R. norvegicus*. **c** Exemplarische Korrelationskurven von *M. musculus* [grün] und *R. norvegicus* [blau] in Form des jeweils ersten und zweiten Chromosoms in einem Diagramm. Der Ausschnitt zeigt eine Vergrößerung eines Teilabschnittes. (Aus: Dehnert et al. (2006).)

Auch wenn hier die Systematik der Korrelationskurven, die eine recht verlässliche Trennung der beiden Spezies erlaubt, im Vordergrund der Betrachtung steht, ist es dennoch interessant, die biologischen Eigenschaften der Ausreißer in Abbildung 2.12 näher zu betrachten. Abbildung 2.13 stellt zwei solche Kenngrößen aller Chromosomen der Maus und der Ratte dar, nämlich den GC-Gehalt (also die Summe der G- und C-Häufigkeiten; dies ist einer der Schlüsselparameter in Diskussionen der Mosaikstruktur einer DNA-Sequenz im Rahmen des Isochoren-Konzepts (Bernardi et al., 1985; Bernardi, 1989, 2000) und lieferte kürzlich Hinweise auf deterministische Prinzipien hinter Aspekten der Genomevolution (Messer et al., 2005)) und die Dichte an CpG-Inseln (also im Wesentlichen die mittlere Zahl homogener Regionen mit einer Erhöhung von CG-Dinukleotiden; für eine statistisch präzise Definition, vgl. auch Takai und Jones (2002)). CpG-Inseln weisen eine deutlich positive Korrelation mit regulatorischen Bereichen von Genen auf (Takai und Jones, 2002). Die Ausreißer in Abbildung 2.12 zeigen sich auch in diesen Eigenschaften als Extremfälle. Aus diesen Beobachtungen, besonders aber aus der Tatsache, dass die weiteren Aspekte des Cluster-Baums sich keinesfalls trivial aus den Eigenschaften aus Abbildung 2.13 ergeben, lässt sich ein schlüssiges Szenario ableiten, das zum Teil die starke Aufmerksamkeit erklärt, die Dinukleotidkenngrößen in der Forschung erfahren haben (Karlin und Ladunga, 1994; Karlin und Mrázek,

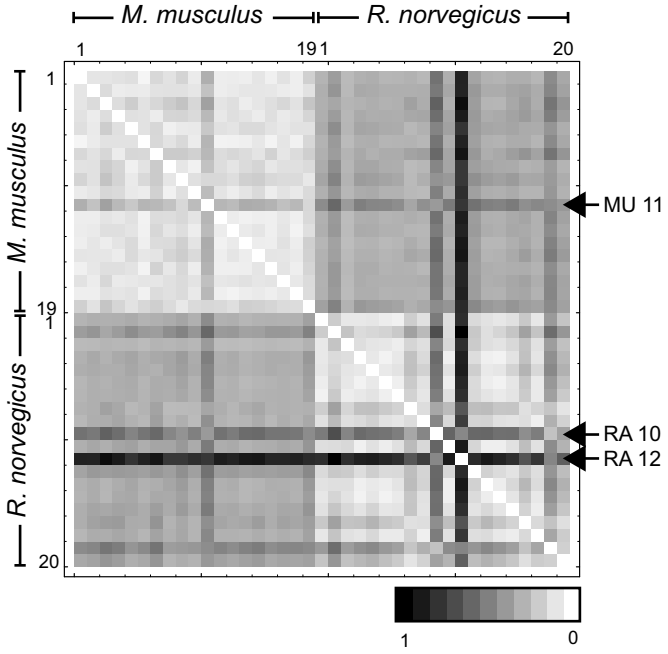

Abb. 2.11. Darstellung der Distanzmatrix in Graustufen für alle Autosomen von *M. musculus* und *R. norvegicus*. Berechnung der paarweisen Distanzen der Korrelationskurven durch die L_1-Norm. Die drei exponierten Chromosomen MU 11, RA 10 und RA 12 sind mit Pfeilen markiert. (Aus: Dehnert et al. (2006).)

1997; Gentles und Karlin, 2001): Extreme Dinukleotidzusammensetzung können gelegentlich alle weitere Information in den statistischen Korrelationen dominieren (oder „überschreiben“), während in allen anderen Fällen die Korrelationsstruktur durch längerreichweitige Eigenschaften als die Dinukleotidebene bestimmt ist.

Betrachten wir nun die Längenabhängigkeit der Clusterung der Chromosomen im *Tree Color Coding* Plot für $p = 50$ in Abbildung 2.14. Beginnend mit den ersten 10 kbp jeder Sequenz beträgt die Schrittweite 10 kbp bis zum Erreichen von 40 Mbp. Für die Sequenzlängen $L = 200$ kbp, $L = 15$ Mbp und $L = 40$ Mbp sind die nach dem TCC-Algorithmus sortierten Bäume explizit dargestellt. Anhand dieser Ausschnitte lässt sich die Clusterung der Bäume sehr gut im Detail studieren und man erhält als Vorbetrachtung auf der methodischen Ebene einen Eindruck, wie der TCC-Algorithmus arbeitet und die Information codiert. Die erste Beobachtung ist, dass die Clusterung der Chromosomen mit größer werdender Sequenzlänge zunimmt. Für eine Länge $L = 200$ kbp zeigt sich noch eine Mixtur von Chromosomen der Maus und Ratte ohne eine Trennung der Spezies. Die Information ist − bedingt durch die kurze Länge − nicht ausreichend, um eine Separation der Spezies zu erzielen. Bei einer Länge der Sequenz von ca. 10 Mbp treten bereits größere Subcluster von Chromosomen jeweils einer Spezies auf, und der Baum bei $L = 15$ Mbp zeigt eine deutliche Unterteilung von Chromosomen der Maus und der Ratte. Ab einer Sequenzlänge von ca. 20 Mbp werden die Chromosomen im TCC-Plot vollständig getrennt bis zum Erreichen der hier betrachteten Maximallänge von 40 Mbp. Betrachtet man nun den der Codierung im TCC-Plot zugrundeliegenden Baum für $L = 40$ Mbp, so wird die Überschätzung der Ordnung des Algorithmus deutlich. Der Baum für $L = 40$ Mpb zeigt eine klare Clusterung der Chromosomen, aber es

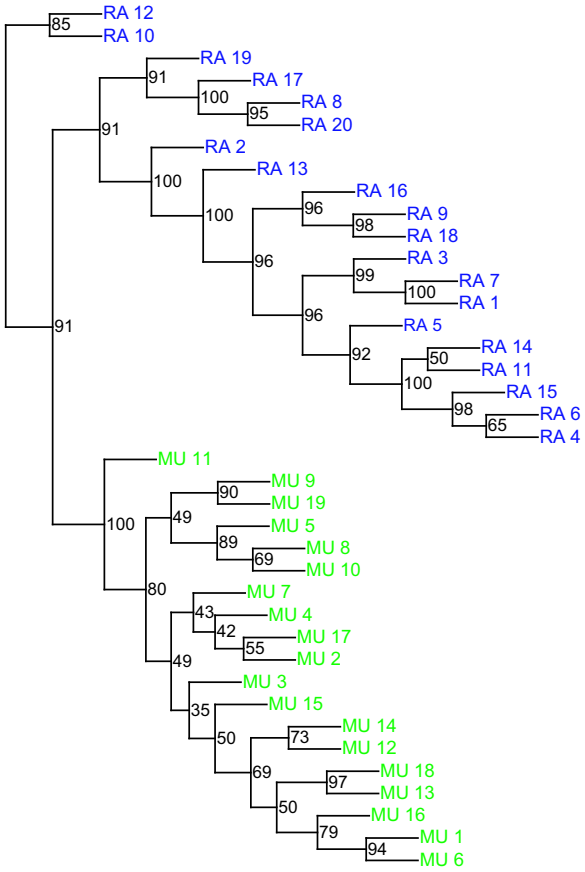

Abb. 2.12. Clusterbaum der Autosomen von *M. musculus* und *R. norvegicus* basierend auf der Distanzmatrix aus Abbildung 2.11. Bootstrap-Werte für 100 Samples sind an den Knoten angegeben. (Aus: Dehnert et al. (2006).)

existiert noch ein relativ großes Cluster von Chromosomen der Ratte, das vor der Maus und den restlichen Chromosomen der Ratte abzweigt. Dieser Baum entspricht nicht dem, den man für die vollen Sequenzlängen erhält.

Auf der inhaltlichen Ebene ist es nun interessant zu untersuchen, wie sich die Längenabhängigkeit der Clusterung unter Variation des Parameters p verhält. Dazu wählt man ein beliebiges aber festes p aus und berechnet den TCC-Plot in Abhängigkeit der Sequenzlängen. In Abbildung 2.15 a ist der TCC für $p = 5$, 20, 30, 50 und 100 dargestellt. Die Erhöhung von p in der Markov-Repräsentation führt zu einer verbesserten Clusterung der Chromosomen. Je größer p gewählt wird, desto geringer ist die benötigte Sequenzlänge, um die Chromosomen im TCC-Plot zu trennen. Für $p = 100$ ist eine Länge von 20 Mbp für die vollständige Aufschlüsselung der Chromosomen ausreichend, während für $p = 20$ eine Sequenzlänge von ca. 40 Mbp benötigt wird. Damit ergibt sich für den Parameter p, der die Größe des betrachteten Symbolabstands festlegt, eine Möglichkeit zur Steuerung des Volumens an abgefragter Information. Im unteren Abschnitt des Bildes ist die gleiche Analyse für die Transinformation dargestellt. Hier zeigt sich erneut der Unterschied zwischen den beiden Repräsentationen. Die Verbesserungen in der Clusterung für längere Sequenzen sind bei

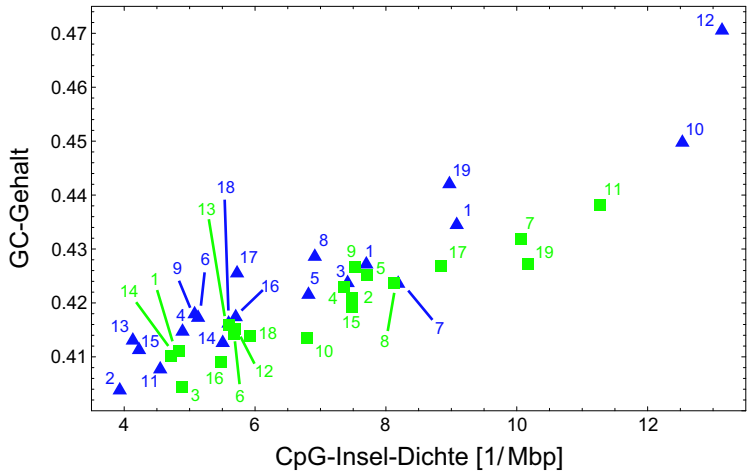

Abb. 2.13. GC-Gehalt vs. Dichte an CpG-Inseln für die Chromosomen von *M. musculus* [grün] und *R. norvegicus* [blau]. Die der Abbildung zugrunde liegenden Daten basieren auf Angaben des *Genome Browser* der *University of California at Santa Cruz* (Hinrichs et al., 2006). (Aus: Dehnert et al. (2006).)

größer werdendem p nur sehr moderat. Für die Transinformation ist eine Trennung der Spezies auch bei hohem p bei der maximalen betrachteten Sequenzlänge von 40 Mbp nicht möglich. In dieser Abbildung wird damit auch deutlich, dass Informationen bezüglich der Spezies nicht nur für kleine Symbolabstände, also zum Beispiel direkt benachbarter Nukleotide existieren, sondern für weit größere Abstände. Die in einem erheblichen Teil der Forschung vorherrschende Ansicht, Speziesinformationen in längerreichweitigen Korrelationen seien ein Epiphänomen der Dinukleotideigenschaften, ist durch diese Ergebnisse widerlegt. Das Messen dieser Abhängigkeiten wird hier mit Hilfe der Markov-Repräsentation erstmals vorgeführt.

Es ist nicht klar, welche Symbolabstände zu einer Verbesserung der Clusterung im TCC-Plot für die Markov-Repräsentation führen, wenn man von $p = 20$ zu $p = 30$ übergeht. Es stellt sich die Frage, wie die Information zur Speziestrennung innerhalb der Korrelationskurven verteilt ist und auf welche Weise diese zunimmt, wenn der Bereich des betrachteten Symbolabstandes vergrößert wird. Zur Beantwortung dieser Fragen benötigt man ein Maß zur Beschreibung des Beitrags einer einzelnen Komponente des Korrelationsvektors zur Trennung zweier Spezies. Der in Kapitel 1.5 eingeführte $|t|$-Wert quantifiziert dies, indem der Abstand zweier Kurvenscharen in jedem Punkt der Korrelationskurve gemessen wird. Dazu wird die Differenz der Mittelwerte der Kurvenscharen in einer Komponente des Korrelationsvektors mit den Varianzen der Kurvenscharen in diesem Punkt normiert. Die Definition des $|t|$-Werts findet sich in Kapitel 1.5 in Gleichung (1.19).

Die Korrelationskurven sind für Maus und Ratte für $p = 30$ zusammen mit dem $|t|$-Wert in Abbildung 2.16 a dargestellt. Der $|t|$-Wert ist klein für Symbolabstände, bei denen die Kurven der Maus und Ratte nahe beieinander liegen oder sich überlagern, und groß in dem Fall, dass es deutliche Unterschiede in den Kurvenscharen gibt. In Abbildung 2.16 a zeigt sich, dass die Speziestrennung das Resultat von Unterschieden ist, die sich über den Korrelationsvektor verteilen. Betrachten wir die Korrelationskurven nun anhand ausgewählter Komponenten etwas genauer, um die Funktionsweise des $|t|$-Wertes besser zu verstehen und damit die Verlässlichkeit des im vorangegangenen

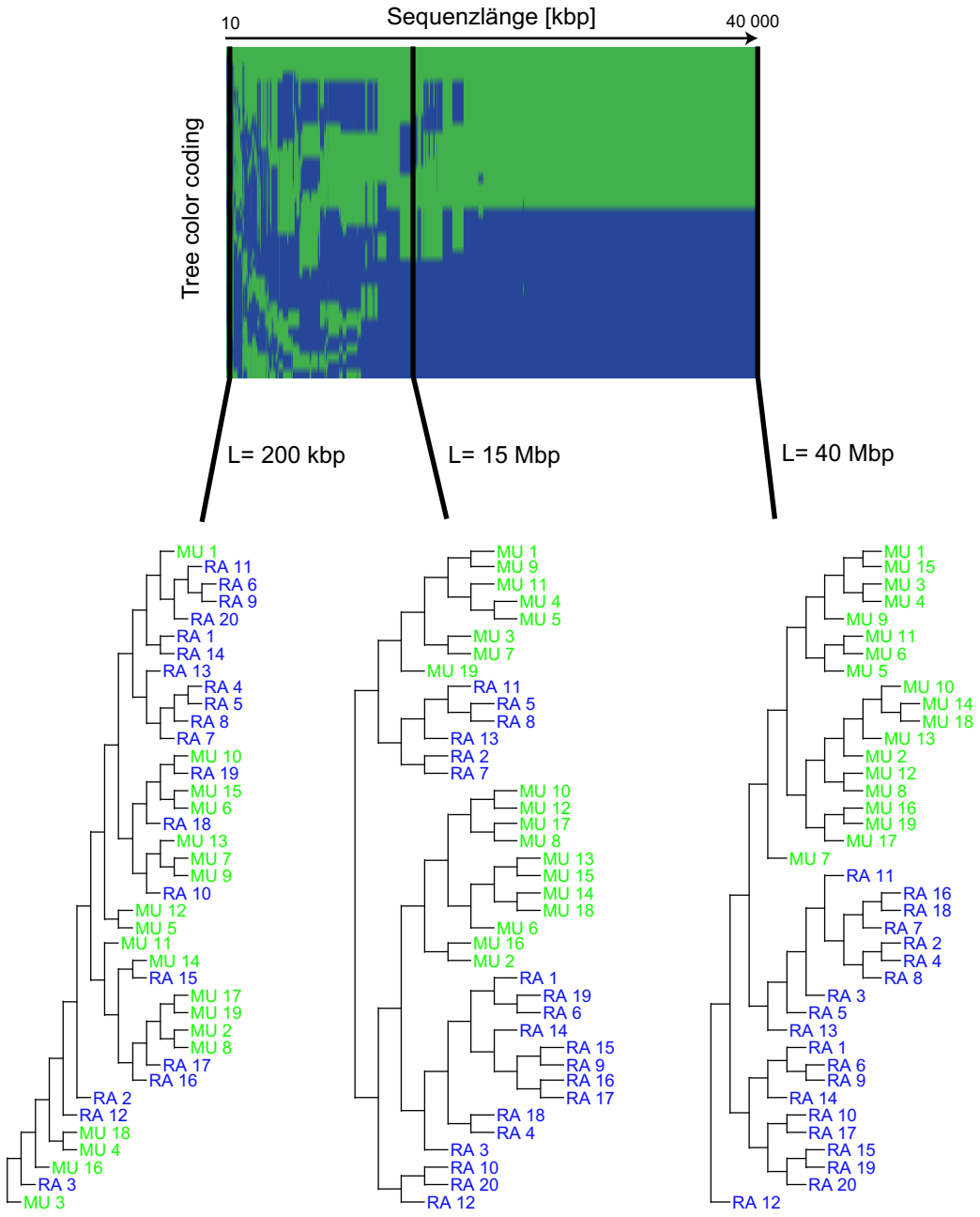

Abb. 2.14. *Tree Color Coding* Plot für die Markov-Repräsentation der Ordnung $p = 50$. Die zugrunde liegende Sequenzlänge wird simultan für alle Chromosomen erhöht, beginnend mit den ersten 10 kbp jeder Sequenz bis zum Erreichen von 40 Mbp. Die Schrittweite beträgt 10 kbp. Wird die maximale Sequenzlänge schon vor den 40 Mbp erreicht, so wird mit der maximal möglichen Sequenzlänge gerechnet. (Aus: Dehnert et al. (2006).)

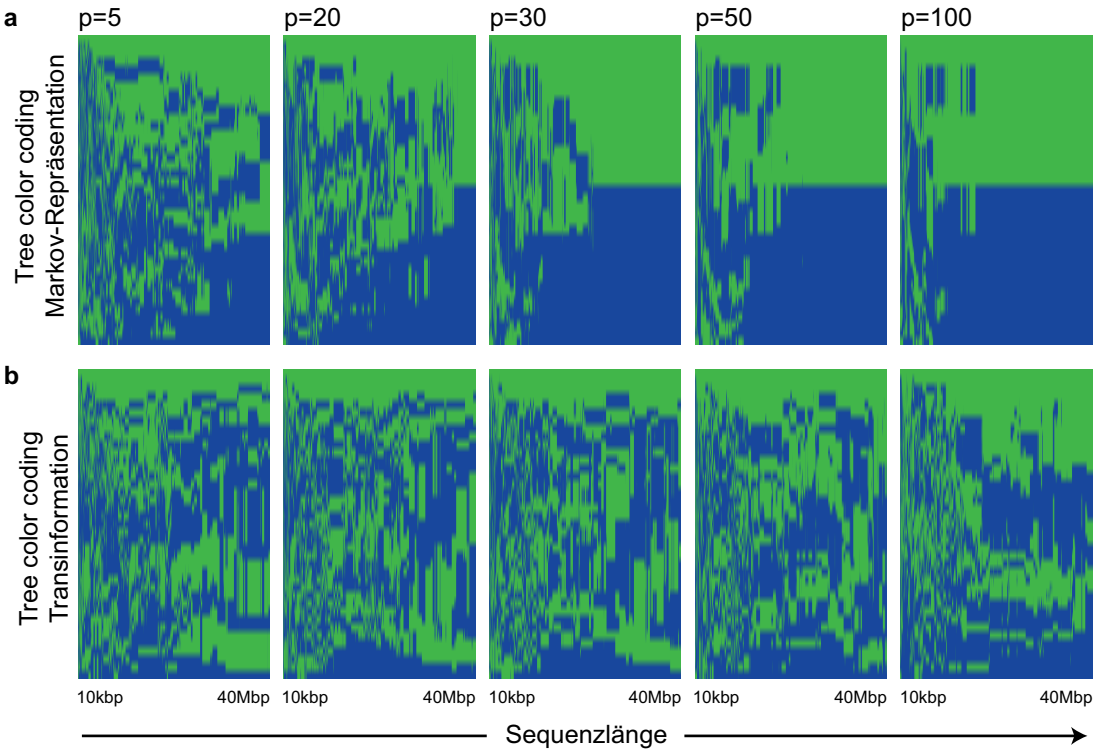

Abb. 2.15. *Tree Color Coding* Plots für **a** Markov-Repräsentation und **b** Transinformation. Es werden unterschiedliche Bereiche von Symbolabständen $k = 1, \ldots, p$ betrachtet. Auf diese Weise wird sowohl die Länge der zugrunde liegenden DNA-Sequenzen als auch die Markov-Ordnung p variiert. (Aus: Dehnert et al. (2006).)

Abschnitt erzielten Ergebnisses einschätzen zu können. Die punktuelle Information, die zur Speziestrennung von Ratte und Maus beiträgt, ist im Symbolabstand 22 am größten. Die Scharen der Korrelationskurven zeigen hier deutliche Unterschiede und eine geringe Varianz, sie sind also stark gebündelt. Auch die Komponente 29 im Korrelationsvektor weist deutliche Unterschiede für Maus und Ratte auf. Die Menge an Information zur Speziestrennung ist aber im Vergleich zur Komponente 22 geringer, da hier eine größere Streuung der Kurven innerhalb einer Spezies beobachtet wird. Als Beispiel für einen kleinen $|t|$-Wert betrachten wir den Symbolabstand 13. Hier überdecken die Korrelationskurven der Ratte die der Maus vollständig, was zu einem geringen Beitrag zur Speziestrennung führt. Auch wenn bei sehr großen oder sehr kleinen Abständen der Scharen von Korrelationskurven der Beitrag zur Speziestrennung in vielen Fällen visuell approximativ möglich ist, so erlaubt der $|t|$-Wert eine quantitative Angabe. Im Abstand $k = 10$ wird deutlich, wie schwierig eine solche Abschätzung anhand rein visueller Anhaltspunkte und bei einer Beschränkung auf nur einen Abstand wäre.

Für alle Kurven lässt sich eine Stabilisierung mit größer werdender Sequenzlänge beobachten. Es zeigt sich außerdem, dass eine Unterscheidung der zwei Wertegruppen nicht in allen Fällen möglich ist. Die Symbolabstände $k = 22$ und $k = 29$ zeigen auch für kurze Sequenzlängen ein hohes Maß an Unterschieden für die beiden Kurvenfamilien. Es stellt sich nun die Frage, ob diese Beobachtung unabhängig von der betrachteten Markov-Ordnung p ist. Dafür sind in Abbildung

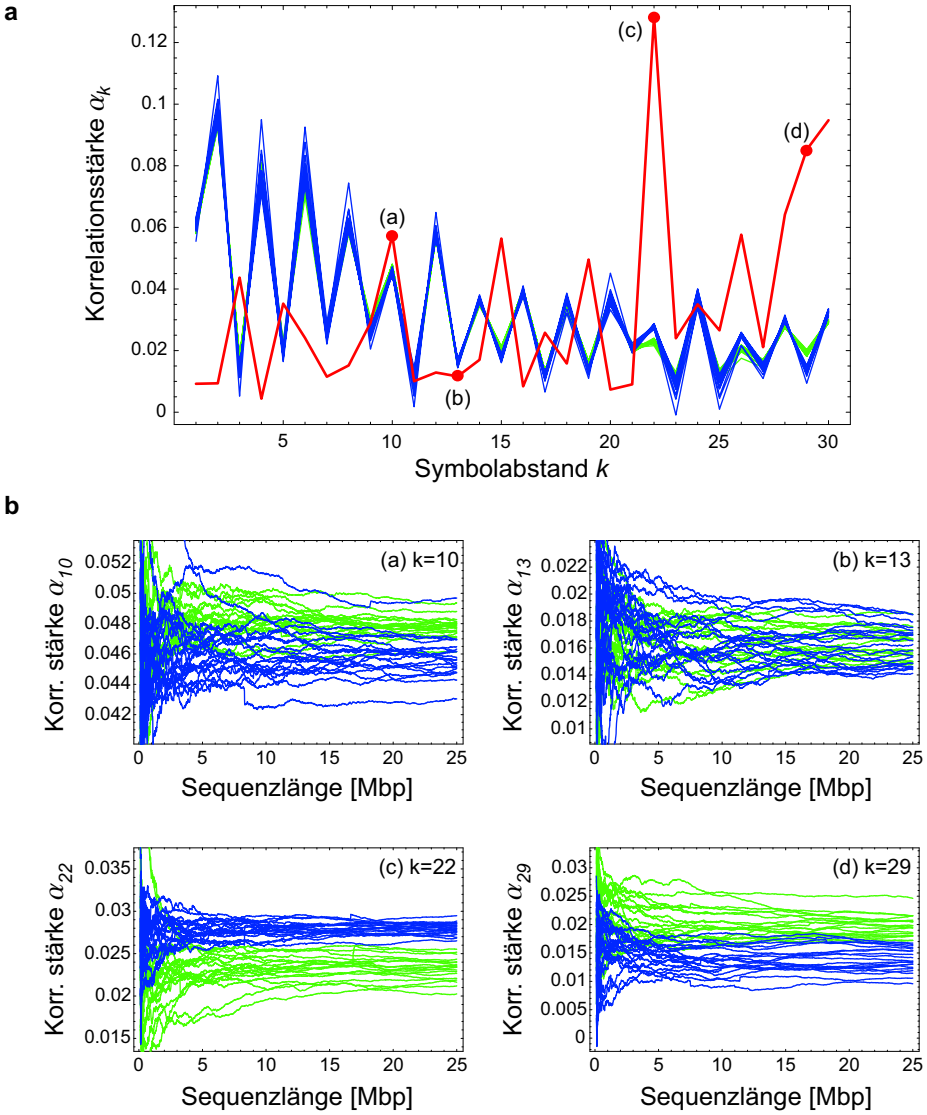

Abb. 2.16. **a** Korrelationskurven der Chromosomen von *M. musculus* [grün, 19 Kurven] und *R. norvegicus* [blau, 20 Kurven] zusammen mit dem $|t|$-Wert als Maß für die Unterschiedlichkeit der beiden Kurvenscharen für jeden Abstand k. Ein hoher bzw. niedriger $|t|$-Wert kennzeichnet einen großen bzw. kleinen Beitrag der Komponente α_k zur Trennung von *M. musculus* und *R. norvegicus*. **b** Abhängigkeit der Korrelationsstärke α_k von der Sequenzlänge der Chromosomen von *M. musculus* und *R. norvegicus* für die Symbolabstände (a) $k = 10$, (b) $k = 13$, (c) $k = 22$ und (d) $k = 29$. (Aus: Dehnert et al. (2005a).)

2.17 die $|t|$-Wert Kurven von $p = 30$ bis $p = 100$ in Schritten von $p = 5$ für die Chromosomen der Maus und Ratte aufgetragen. In diesem Fall ist der $|t|$-Wert jedoch nicht auf die Summe von Eins normiert. Man sieht eine sehr starke Ähnlichkeit der Kurven, wenn auch keine vollständige Überdeckung. Der Parametervektor $\vec{\alpha}$ des DAR(p)-Prozesses ist nicht − wie die Transinformation − unabhängig von $p = k_{max}$. Diese Eigenschaft der Markov-Repräsentation wird in Kapitel 1 aus-

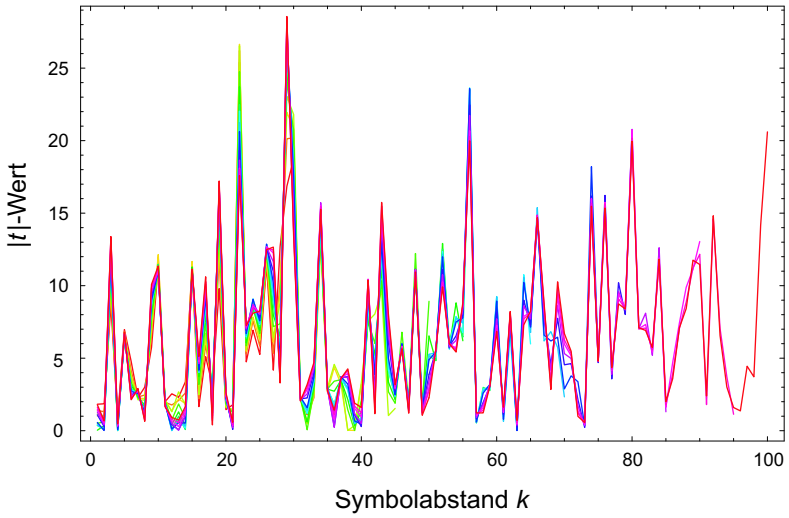

Abb. 2.17. $|t|$-Wert Kurven (ohne Normierung auf die Summe von Eins) für unterschiedliche Ordnungen p der Markov-Repräsentation für die Chromosomen von *M. musculus* und *R. norvegicus*.

führlich diskutiert. Obwohl die Normierungseigenschaft der Yule-Walker-Gleichungen sich in den Korrelationskurven widerspiegelt, zeigt diese Analyse aber ganz deutlich, dass die Schwankungen in Abhängigkeit von p sehr gering sind, und dass Symbolabstände k, die einen großen $|t|$-Wert zeigen, diesen unabhängig von p aufweisen. Außerdem wird anhand dieser Abbildung noch einmal sehr deutlich, dass eine Vergrößerung des betrachteten Bereichs von Symbolabständen zu mehr speziestrennender Information führt. Dies ist wie bereits diskutiert ein starkes Argument dafür, dass Speziesunterschiede weit über die unterschiedlichen Verteilungen von Dinukleotidhäufigkeiten hinausgehen.

2.3 Schimpanse und Huhn

Die im Vorangegangenen diskutierten Analysen werden nun erneut um mehrere Spezies erweitert. Dabei sind *Gallus gallus* (Huhn) und *Pan troglodytes* (Schimpanse) von besonderem Interesse. In Abbildung 2.18 sind die Korrelationskurven dieser Spezies zusammen mit denen von Mensch und Maus (ohne Berücksichtigung der Geschlechtschromosomen) für p=30 aufgetragen.[3] In dieser Darstellung zeigt sich noch einmal der Sachverhalt, dass nah verwandte Spezies ähnliche Korrelationskurven zeigen, wie im Fall von Mensch und Schimpanse deutlich zu sehen ist. Abbildung 2.18 b mit den Chromosomen des Menschen lässt sich dabei visuell nicht von Abbildung 2.18 a unterscheiden, in der die Korrelationskurven des Schimpansen für alle 23 Autosomen aufgetragen sind. Die Korrelationskurven des Huhns (Abbildung 2.18 c) zeigen deutliche Unterschiede zu den Kurven von Mensch und Maus. Über die Bestimmung der Distanzmatrix durch Bilden der paarweisen Abstände mit Hilfe der L_1-Norm gelangt man nach Anwendung des UPGMA-Algorithmus zu dem in Abbildung 2.19 dargestellten Clusterbaum. Die Zahlen an den Knoten stellen Bootstrap-Werte dar. Zwei Eindrücke beherrschen das Bild: Die Chromosomen des Men-

[3] Das Chromosom 16 von *G. gallus* wird in dieser und den folgenden Analysen nicht berücksichtigt, da mehr als 20% der Sequenz aus nicht identifizierten Nukleotiden bestehen.

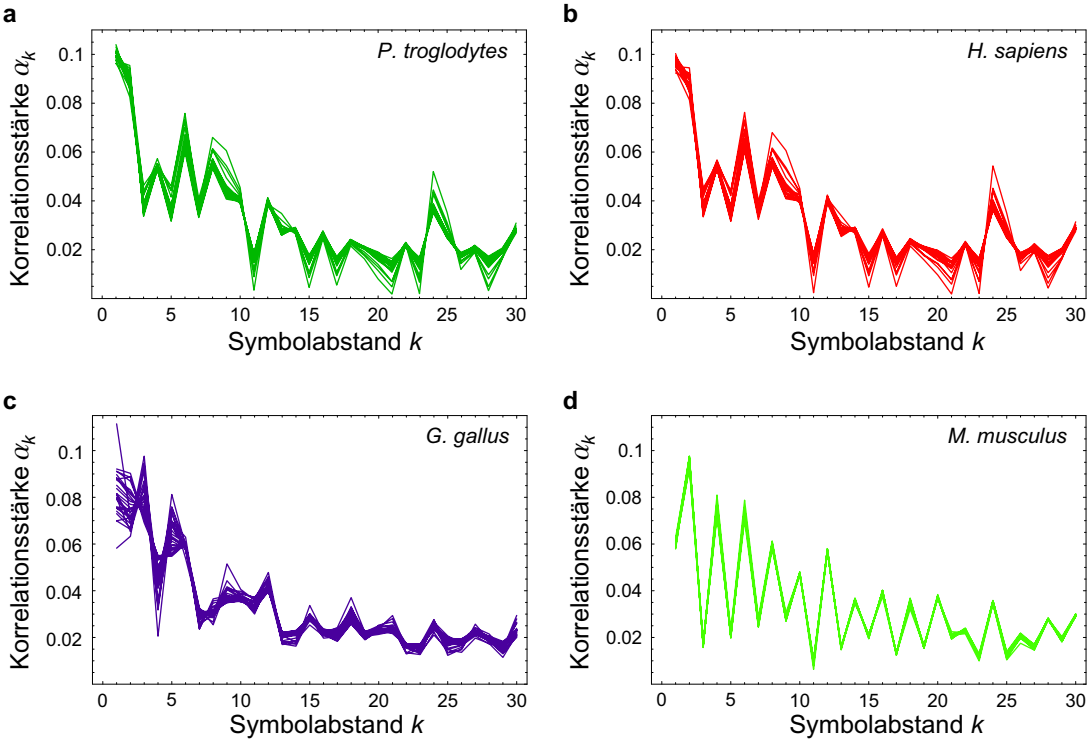

Abb. 2.18. Korrelationskurven der Markov-Repräsentation für die Chromosomen der folgenden Spezies: **a** *P. troglodytes* [23 Kurven], **b** *H. sapiens* [22 Kurven], **c** *G. gallus* [27 Kurven] und **d** *M. musculus* [19 Kurven]. (Angepasst aus: Dehnert et al. (2005a).)

schen und des Schimpansen durchmischen sich, und die Chromosomen des Huhns werden neben dem Cluster aus Chromosomen des Menschen und des Schimpansen eingeordnet, noch vor der Abzweigung von Maus und Ratte. Für die Chromosomen des Huhns ergibt sich damit eine aus phylogenetischer Sicht falsche Einordnung im Baum (zur Übersicht siehe Hedges (2002)). Die Hoffnung ist, dass die falsche Einordnung Aufschluss über biologische Ursachen der unterschiedlichen Interspezies-Signaturen und die Intraspezies-Synchronisation geben kann. Darauf wird im späteren Verlauf der Arbeit näher eingegangen.

Betrachten wir zunächst die Clusterung der Chromosomen des Menschen und des Schimpansen. Die Chromosomen dieser Spezies werden im Baum nicht getrennt, sondern es zeigt sich eine Mixtur mit nur wenigen kleinen Subclustern von Chromosomen einer Spezies. Man beobachtet stattdessen eine größere Anzahl von Paarbildungen von Chromosomen des Menschen und des Schimpansen, die hohe Bootstrap-Werte aufweisen, was auf eine robuste Clusterung hindeutet. So zeigen die Paare HU 19/CH 20, HU 17/CH 19, HU 22/CH 23, HU 16/CH 18, HU 20/CH 21 und HU 1/CH 1 Bootstrap-Werte von 100. All diese Paare sind orthologe Chromosomen des Menschen und Schimpansen. Bei dem anderen eng verwandten Paar von Spezies in dieser Analyse, nämlich Maus und Ratte, wird keine paarweise Clusterung von Chromosomen beobachtet. Das System zur Bezeichnung der Chromosomen des Schimpansen ist kürzlich auf Vorschlag von McConkey (2004) erneuert worden. Die Umbenennung der Chromosomen des Schimpansen erfolgt dabei so, dass die Bezeichnungen dieser Chromosomen mit den orthologen Chromosomen des Menschen

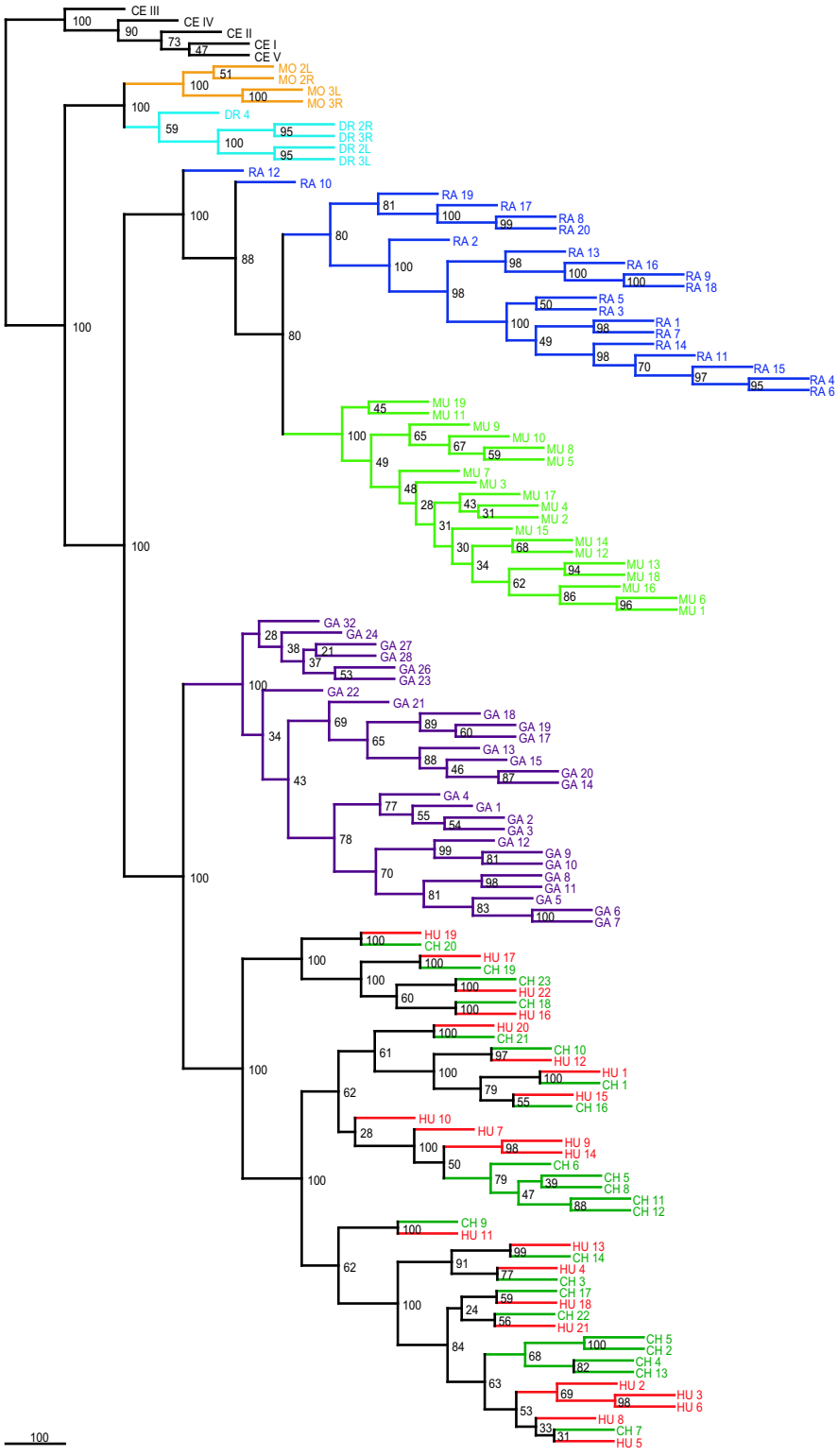

Abb. 2.19. Clusterbaum für 125 Chromosomen von 8 eukaryotischen Spezies. Die Abbildung zeigt den *Consensus*-Baum von 100 Bootstrap-Samples mit einer Angabe der Bootstrap-Werte an den Knoten der Zweige. (Aus: Dehnert et al. (2005a).)

korrespondieren. Die neue Nomenklatur wurde bereits von NCBI[4] und Ensembl[5] übernommen, der *Genome Browser*[6] nutzt noch die originalen Benennungen, die auch hier gewählt wurden. In Tabelle C.5 in Anhang C wird die alte und neue schimpansische Chromosomenbezeichnung der menschlichen Einteilung gegenübergestellt.

In Abbildung 2.20 wird die Längenabhängigkeit der Clusterung mit Hilfe des TCC-Plots untersucht. Es zeigt sich deutlich die Durchmischung der Chromosomen für Mensch und Schimpanse für alle betrachteten Sequenzlängen, sowie die Trennung der Chromosomen für wachsende Sequenzlängen bei allen anderen in der Analyse betrachteten Spezies. Die Position der Chromosomen des Huhns festigt sich dabei bereits bei relativ kurzen Sequenzlängen (ca. 100 kbp) und verbleibt in der Position neben der Gruppe von Chromosomen des Menschen und Schimpansen bis zum Erreichen der maximalen Sequenzlänge von 25 Mbp.

Die Korrelationskurven mit $p = 30$ von Mensch und Schimpanse erlauben keine Trennung der Spezies. Die Untersuchung bei Maus und Ratte hat gezeigt, dass mit wachsendem p die Trennung der Spezies im TCC-Plot immer kürzere Sequenzen benötigt. Es liegt deshalb die Frage nahe, ob die Trennung der Chromosomen bei größerem p möglich wird. In Abbildung 2.21 sind Korrelationskurven des Menschen und des Schimpansen bis $p = 300$ aufgezeichnet. Die Korrelationskurven der Chromosomen liegen übereinander und zeigen zwischen den Spezies keine systematischen Unterschiede. Das Resultat einer Clusteranalyse ausschließlich dieser beiden Spezies zeigt, dass auch für $p = 300$ die Chromosomen nicht getrennt werden können. Der Eindruck, dass die Korrelationskurven von Mensch und Schimpanse keine Trennung der Spezies erlauben, bestätigt sich damit. Die Verzweigungen im Baum in Abbildung 2.19 bleiben darüber hinaus erhalten und dokumentieren damit die Robustheit des Baums. Auf Basis der im Rahmen dieser Arbeit diskutierten Methode der Korrelationskurven lassen sich die Chromosomen der Spezies Mensch und Schimpanse nicht trennen. Stattdessen findet bzw. bestätigt die Methode die Homologien zwischen Mensch und Schimpanse.

Betrachten wir nun die Korrelationskurven des Huhns. In Abbildung 2.22 sind diese zusammen mit denen des Menschen und den sich ergebenden $|t|$-Werten aufgetragen. Die erste Erkenntnis bei Betrachtung dieses Bildes ist, dass ein nahes Beieinanderliegen von Clustern in Abbildung 2.19 nicht automatisch mit einer hohen Ähnlichkeit der zugrunde liegenden Korrelationskurven einhergeht. Die Schar der Korrelationskurven des Menschen zeigt klare und systematische Unterschiede zu der Familie der Korrelationskurven des Huhns. Der $|t|$-Wert als Funktion des Abstandes zeigt klare Peaks für solche Abstände von k, denen man auf rein visueller Basis den größten Beitrag zur Speziesunterscheidung zuweisen würde. Auch hier muss untersucht werden, ob eine Erweiterung des betrachteten Symbolabstands, etwa auf $p = 100$, zu einer Änderung, in diesem Fall der Position der Chromosomen des Huhns, im Vergleich zum Clusterbaum in Abbildung 2.19 führt. Der Clusterbaum für Korrelationskurven bis $p = 100$ zeigt jedoch die gleiche Metastruktur wie für $p = 30$, und dabei bleibt insbesondere die Platzierung der Chromosomen des Huhns erhalten. Damit wird deutlich, dass diese Einordnung kein Artefakt der Wahl der Parameter der Analyse darstellt, mit deren Hilfe das Volumen der abgefragten Information gesteuert werden kann.

Es stellt sich die generelle Frage, welche Komponenten innerhalb der DNA-Sequenzen für die beobachtete Signatur verantwortlich sind und ob man anhand der aus phylogenetischer Sicht falschen

[4] http://www.ncbi.nlm.nih.gov/
[5] http://www.ensembl.org/
[6] http://genome.ucsc.edu/

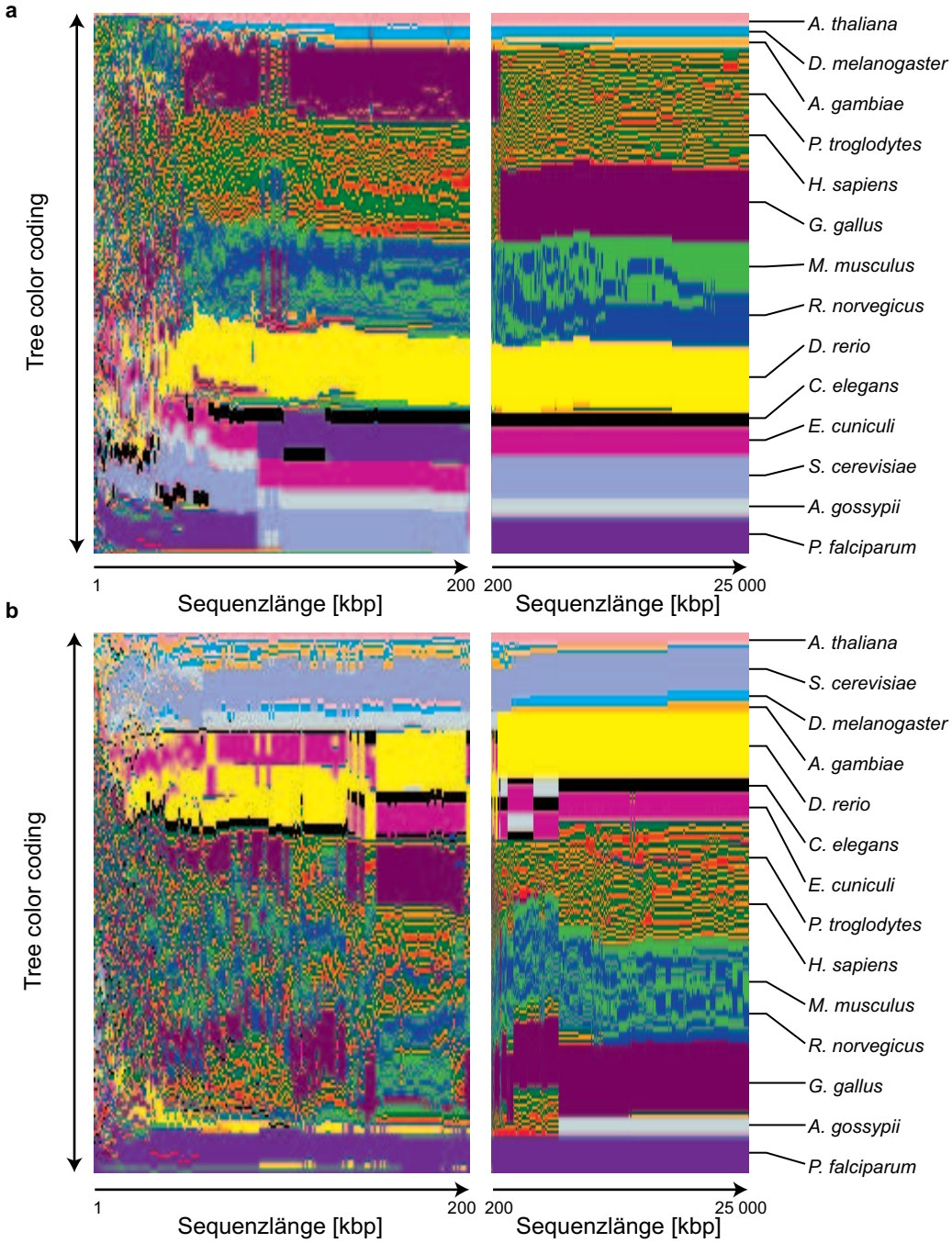

Abb. 2.20. *Tree Color Coding* Plot für 14 eukaryotische Spezies für **a** die DAR(p) und **b** die $I(k)$ Repräsentation der Korrelationskurven. Die Länge der zugrundeliegenden DNA-Sequenzen wird variiert. Für jede Länge wird ein Clusterbaum erstellt, der dann mit Hilfe des TCC-Algorithmus in eine Abfolge von Farbsegmenten übersetzt wird. Beginnend mit den ersten 1000 Basen jeder Sequenz werden alle Sequenzen der 203 Chromosomen simultan mit einer Schrittweite von 1 kbp (bis 200 kbp) und von 10 kbp (bis 25 Mbp) erhöht. Für den Fall, dass die Sequenzlänge eines Chromosoms kürzer als 25 Mbp ist, wird die Sequenz bei maximaler Länge konstant gehalten. (Angepasst aus: Dehnert et al. (2005a).)

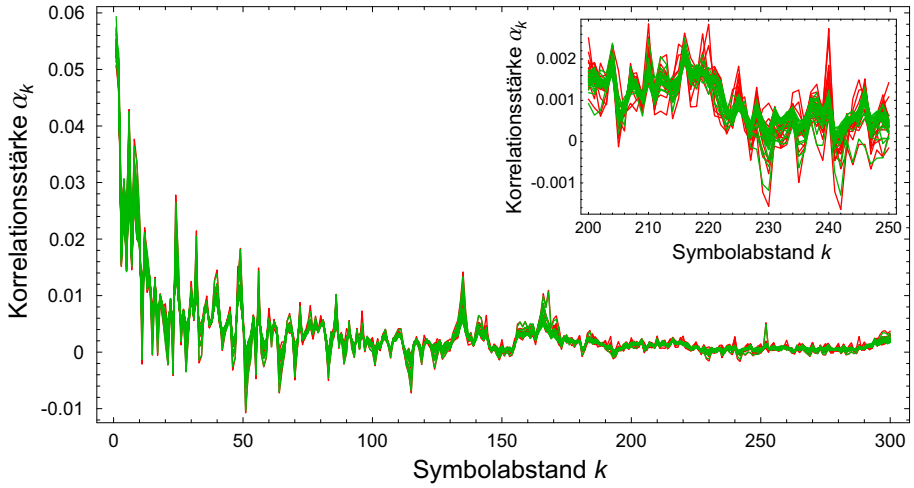

Abb. 2.21. Korrelationskurven von *H. sapiens* [rot, 22 Kurven] und *P. troglodytes* [grün, 23 Kurven] der Markov-Repräsentation für $p = 300$. Der Ausschnitt zeigt die Korrelationsstärke für $k = 200, \ldots, 250$.

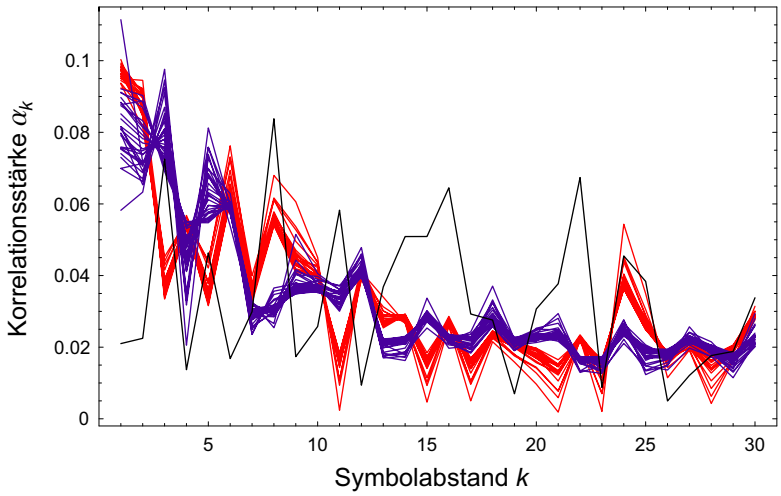

Abb. 2.22. Korrelationskurven für die Chromosomen von *H. sapiens* [rot, 22 Kurven] und von *G. gallus* [violett, 27 Kurven], zusammen mit dem $|t|$-Wert für die Kurvenscharen. (Aus: Dehnert et al. (2005a).)

Einordnung des Huhns einen Anhaltspunkt dafür findet. Dieser Frage soll im folgenden Kapitel nachgegangen werden.

2.4 Biologische Ursachen statistischer Korrelationen in DNA-Sequenzen

In der folgenden kurzen Übersicht orientiere ich mich erneut an Hütt und Dehnert (2006). Aus biologischer Sicht unterscheiden sich Spezies auf ganz unterschiedlichen Ebenen. Sie lassen sich zum

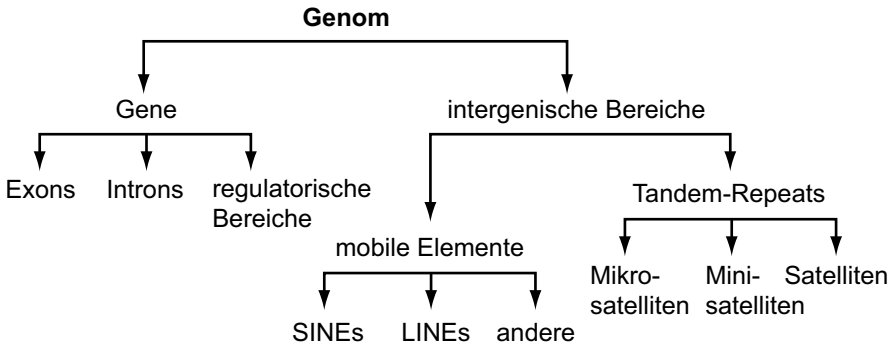

Abb. 2.23. Hierarchie von Beiträgen zu genomweiten eukaryotischen DNA-Sequenzen. (Aus: Hütt und Dehnert (2006).)

Beispiel auf Basis von morphologischen und phänotypischen Merkmalen einteilen. Die 16S bzw. 18S rRNA Analyse stellt auf der anderen Seite das Standardwerkzeug bei einer genotypischen Unterscheidung dar. Die statistischen Merkmale, die in dieser Arbeit analysiert werden, sind das Resultat verschiedener Sequenzkomponenten, aus denen sich eukaryotische Genome aufbauen. Eukaryotische Genome sind eine Vermengung codierender und nicht-codierender Sequenzsegmente, in der wiederum die codierenden Bereiche systematisch von nicht-translatierten Regionen durchsetzt sind. Typische Bestandteile der Gene sind *Exons*, *Introns* und regulatorische Elemente wie Promotorregionen und *Enhancer* oder *Silencer*. In den intergenischen Bereichen finden sich *Pseudogene*, also Genen ähnliche Strukturen, die von der zellulären Maschinerie nicht mehr abgelesen werden, und regulatorische Bereiche, die auf (meist nahegelegene) Gene wirken. Vor allem aber sind diese intergenischen Regionen geprägt von dynamischen Prozessen auf einer evolutionären Zeitskala. In diesen Prozessen werden einzelne Nukleotide oder Nukleotidgruppen lokal vervielfältigt oder ganze größere Segmente ausgeschnitten und an anderer Stelle wieder eingesetzt. In diesen Bereichen wird zwischen *mobilen Elementen* und *Tandem-Repeats* unterschieden. Beide Gruppen repetitiver Elemente stellen in vielen eukaryotischen Genomen einen erheblichen Anteil am Genom dar (Human Genome Sequencing Consortium, 2001; Mouse Genome Sequencing Consortium, 2002; Rat Genome Sequencing Project Consortium, 2004; The Chimpanzee Sequencing and Analysis Consortium, 2005). Abbildung 2.23 zeigt einige zentrale Elemente des Genominventars, wobei dies eine stark vereinfachte Sichtweise darstellt.

Betrachten wir nun die intergenischen Bereiche, u.a. bestehend aus mobilen Elementen und Tandem-Repeats etwas genauer. Mobile Elemente sind DNA-Sequenzen, die die Fähigkeit haben, sich in ihrer Ursprungszelle aus dem Genom herauszulösen und an anderer Stelle in das Genom einzufügen (zur Übersicht siehe z.B. Luning Prak und Kazazian (2000); Deininger und Batzer (2002)). Zu solchen mobilen Elementen gehören *DNA-Transposons* und *Retrotransposons*. DNA-Transposons werden in der Regel aus dem Genom entfernt und an einer anderen Stelle wieder eingesetzt (*cut-and-paste*). Retrotransposons dagegen werden in RNA transkribiert, danach durch die reverse Transkriptase wieder in DNA übersetzt und dann in das Genom integriert (*copy-and-paste*). Aufgrund ihrer Bedeutung für die Genom-Evolution sind Retrotransposons von großem Interesse (Batzer und Deininger, 2002; Deininger et al., 2003; Kazazian, 2004; Hedges und Batzer, 2005). Sie untergliedern sich unter anderem in kurze und lange Elemente: *short interspersed elements*, SINEs, und *long interspersed elements*, LINEs. Im menschlichen Genom stellen *Alu-*

Repeats die wichtigste Klasse von SINEs und L1-Repeats die wichtigste Klasse von LINEs dar (Human Genome Sequencing Consortium, 2001). Neben diesen LINEs und SINEs gibt es noch Retrotransposons, die durch Repeatregionen in den Endbereichen (*long terminal repeats, LTRs*) gekennzeichnet sind. Neben den mobilen Elementen werden in Abbildung 2.23 Tandem-Repeats als Beitrag zu den intergenischen Bereichen genannt. Damit sind Regionen gemeint, die im Wesentlichen aus vielen Wiederholungen eines bestimmten kurzen Segments bestehen. Je nach Länge des wiederholten Segments unterscheidet man Satelliten, Mini- und Mikrosatelliten (Ellegren, 2004), wobei den Mikrosatelliten eine besondere Bedeutung zukommt, da ihre evolutionäre Vervielfältigung ein direktes Produkt bestimmter Aspekte der Replikation darstellt. Mikrosatelliten treten vorwiegend in nicht-codierender DNA auf. Ihr wichtigster Vervielfältigungsmechanismus sind Paarungsfehler durch Schleifenbildung bei der DNA-Replikation und − eng damit verknüpft − Positionsfehler der DNA-Polymerase (Kunkel und Bebenek, 2000; Ellegren, 2004). Die starke Reduktion solcher Mikrosatelliten in codierenden Sequenzbereichen wird letztlich durch Reparaturmechanismen erreicht, mit denen Mutationen verhindert werden, die den Leserahmen gefährden würden (engl. *frame shift mutations*). Eine Ausnahme bilden Trinukleotid-Repeats, die von der Reparaturmaschinerie schwerer identifiziert werden können und als Gendefekte Ursache bestimmter genetischer Krankheiten sind.

Ein möglicher Ansatz zur Erklärung einiger Befunde aus Kapitel 2.3 ist, dass durch ein schrittweises Eliminieren einzelner funktionell benennbarer Bestandteile der Sequenz sich entsprechende Beiträge zur beobachteten Synchronisation und der speziesspezifischen Information der Korrelationsstruktur messen lassen. Letztendlich ist dabei das Ziel, die biologischen Prozesse, die für die beobachtete Synchronisation verantwortlich sind, zu identifizieren. Ungefähr 60 % der proteincodierenden Gene des Huhns haben ein einzelnes menschliches orthologes Gen (International Chicken Genome Sequencing Consortium, 2004). Dabei weisen diese Gene, bedingt durch den hohen Selektionsdruck, eine hohe Konservierung in den Exons auf und nur geringe Ähnlichkeiten in den nicht-codierenden Introns. Insgesamt stellen codierende Regionen im Genom des Huhns jedoch nur 4% der Sequenz dar. Im Vergleich zu anderen sequenzierten Wirbeltieren wurde im Genom des Huhns nur eine geringe Menge von repetitiven Elementen annotiert. Bei Säugetieren liegt der Anteil dieser Elemente zwischen 40-50% des gesamten Genoms, während bei dem Huhn nur 11% der Sequenz dieser Klasse zugeordnet werden. Eine in allen bisher sequenzierten Spezies aufgefundene Art von repetitiven Elementen, die short interspersed elements (SINEs), sind im Genom des Huhns seit 50 Millionen Jahren nicht mehr aktiv, d.h. sie vermehren sich nicht mehr im Genom und sind damit fast gänzlich verschwunden (International Chicken Genome Sequencing Consortium, 2004).

2.4.1 Maskierung von Genen

Eine Aufteilung in Gene und intergenische Bereiche, wie in Abbildung 2.23 dargestellt, dient als erste Unterteilung von biologisch motivierten Komponenten. Der von der *University of California at Santa Cruz* (UCSC) betriebene *Genome Browser* (Hinrichs et al., 2006) ist eine Metadatenbank, in der unter anderem die Annotation von Genen, regulatorischen Bereichen und repetitiven Elementen auf der Ebene ganzer Chromosomen abrufbar ist. Mit Hilfe eines Menüs lassen sich diese Annotationen auswählen und als ASCII-Datei lokal speichern. Die Positionsangabe der Elemente erlaubt nun eine Maskierung dieser Sequenzabschnitte innerhalb der einzelnen Chromosomen.

Die Maskierung durch Ausschneiden der als in der Datenbank RefSeq[7] am NCBI annotierten Gene in den Sequenzen führt in den meisten Fällen zu keiner merklichen Änderung in den Korrelationskurven. Einzig Spezies, die einen großen Anteil codierender DNA besitzen (unter den hier betrachteten Datensätzen also vor allem *C. elegans*, Drosophila und Moskito), weisen deutliche systematische Änderungen auf. Hier ist die eingangs erwähnte charakteristische Periode-3-Oszillation codierender Bereiche auch schon in den ursprünglichen Korrelationskurven recht klar sichtbar (vgl. Abbildung B.1 in Anhang B). Auch nach der Maskierung bleibt die Speziesidentität jedoch erhalten: Es liegt weiterhin eine hohe Synchronisation der Korrelationskurven innerhalb einer jeden Spezies vor, und die jeweiligen Kurvenscharen unterscheiden sich systematisch.

2.4.2 Maskierung von repetitiven Elementen

Es existiert eine größere Anzahl unterschiedlicher Softwarepakete mit deren Hilfe repetitive Elemente in DNA-Sequenzen detektiert werden können. Im Fall von Tandem-Repeats basieren diese Programme oft auf Algorithmen die nach rein mathematischen Gesichtspunkten operieren (Benson, 1999; Castelo et al., 2002). Mobile Elemente werden dagegen vornehmlich durch den Abgleich mit speziellen Datenbanken identifiziert. Eine solche Datenbank ist *Repbase* (Jurka et al., 2005), die umfassende speziesspezifische Sammlungen von bekannten mobilen Elementen und Tandem-Repeats zur Verfügung stellt. Die Programme *Repeatmasker* (Smit et al., 2004) und *CENSOR* (Jurka et al., 1996) erlauben die Lokalisation und Klassifizierung von repetitiver DNA auf Basis dieser Datenbank. Die nach den Resultaten des *Repeatmasker* vorgenommenen Annotationen repetitiver Elemente eukaryotischer Genome sind im *Genome Browser* abrufbar.

Abbildung 2.24 zeigt die Korrelationskurven für acht eukaryotische Spezies nach Überschreiben aller durch Repeatmasker annotierten Repeats mit zufälligen Symbolsequenzen. Die Verteilung der Wahrscheinlichkeiten der einzelnen Basen in diesen Abschnitten entspricht der Verteilung in der jeweiligen unmaskierten Sequenz. Es wird also für jedes Chromosom eine individuelle Anpassung der zufälligen Sequenzen vorgenommen. Um den Einfluss dieser Art der Maskierung jenem durch das Ausschneiden repetitiver Elemente aus der Sequenz gegenüberzustellen, sind in Anhang B in Abbildung B.2 die Korrelationskurven für diese zweite Maskierungsart angegeben.

Für den hier betrachteten Datensatz von *C. elegans* werden 17% der sequenzierten DNA-Sequenz der Autosomen als repetitiv ausgewiesen. Die folgenden Angaben beziehen sich immer auf die Menge an repetitiven Elementen, bezogen auf die Größe der annotierten Autosomen im jeweiligen Genom. Die Maskierung dieser Abschnitte führt bei *C. elegans* auf eine geringe Änderung in den Korrelationskurven, wie in Abbildung 2.24 a zu sehen ist. Die zum Vergleich angegebenen Kurven der originalen Chromosomen liegen dicht neben denen der maskierten. Die optisch größte Änderung ergibt sich für den Symbolabstand $k = 1$, also für direkt benachbarte Basen. Die maskierten Sequenzen weisen eine etwas geringere Korrelationsstärke in diesem Abstand auf. Ein ähnliches Bild ergibt sich für die maskierten Chromosomen von Huhn, Moskito und Drosophila (Abbildung 2.24 b, c, und d). Trotz leicht unterschiedlicher Verläufe zwischen maskierten und unmaskierten Chromosomen bleibt die Signatur der Spezies auch hier vollständig erhalten. Die Abweichungen zwischen den jeweiligen Kurvenscharen sind gering und nicht systematisch. Der prozentuale Anteil von überschriebenen Repeats beträgt in allen drei Genomen ca. 9%. Ein ganz

[7] http://www.ncbi.nlm.nih.gov/RefSeq/

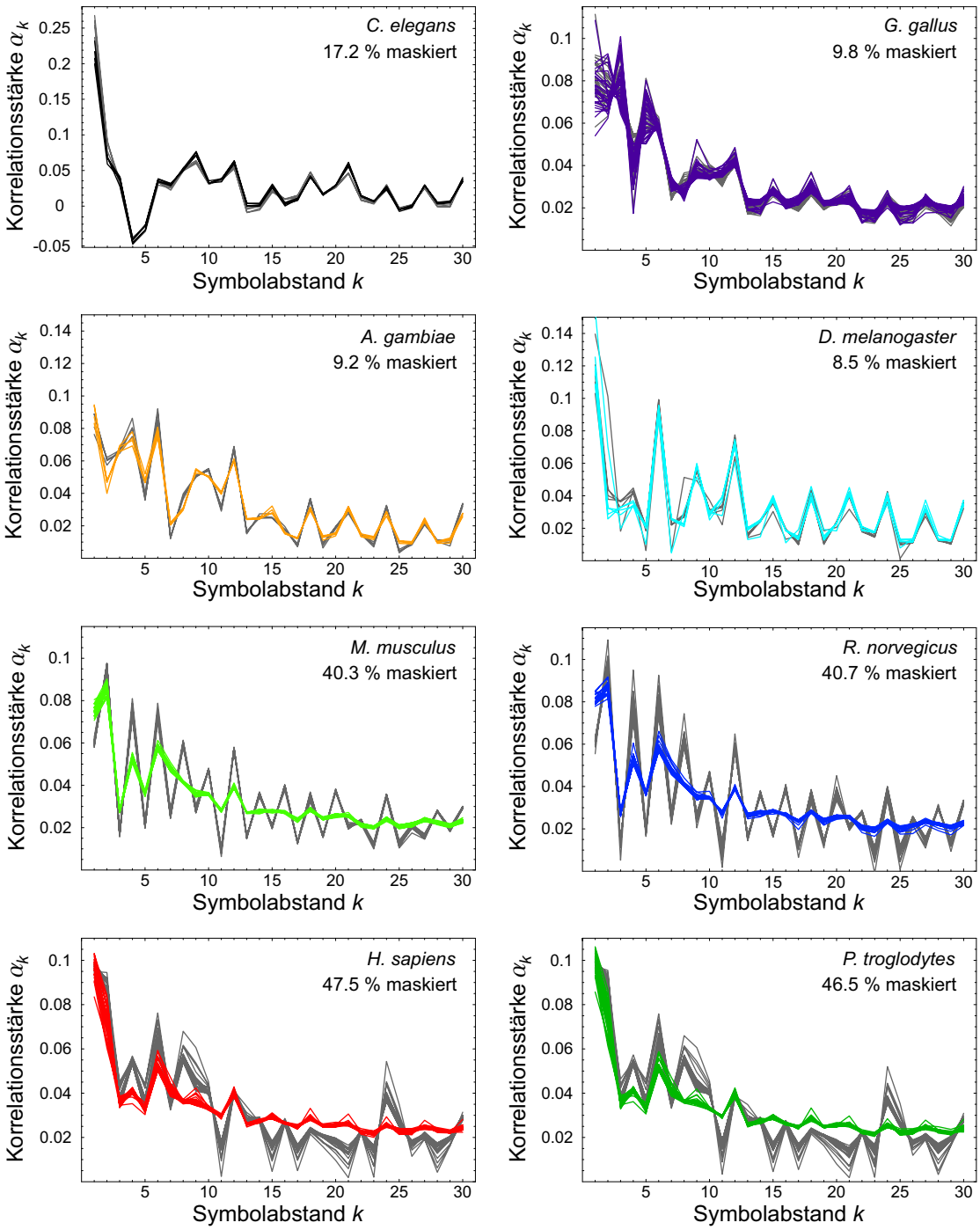

Abb. 2.24. Korrelationskurven nach der Maskierung der repetitiven Elemente [farbig] für die Chromosomen der in der Abbildung genannten Spezies mit Angabe des prozentualen Anteils maskierter Basen, im Vergleich zu den Korrelationskurven der unmaskierten Chromosomen [grau].

anderes Bild ergibt sich für die vier anderen in der Analyse verbliebenen Spezies. Die Korrelationskurven der Maus zeigen nach der Maskierung einen deutlich flacheren Verlauf (Abbildung 2.24 e). Die mittlere Amplitude der Kurven der maskierten Chromosomen sinkt deutlich und es ergeben sich sehr gut sichtbare Unterschiede für alle Symbolabstände. Auch die Position von Maxima und Minima ist nicht in allen Fällen identisch, z.B. fällt für die Abstände 21 und 27 ein Maximum (maskierte Sequenz) auf ein Minimum (originale Sequenz). Die Sequenzmenge der zu den Repeats zählenden Komponenten beträgt bei der Maus mit ca. 40% genauso viel wie bei der Ratte. Die Korrelationskurven der maskierten und orginalen Chromosomen der Ratte zeigen ähnliche Unterschiede wie die der Maus. Die Varianz ist in den maskierten und unmaskierten Kurvenscharen der Ratte ungefähr gleich und etwas größer als bei der Maus. Die Korrelationskurven der maskierten Chromosomen des Menschen und des Schimpansen (Abbildung 2.24 g und h) verhalten sich in diesem Punkt anders. Auch hier ist ein deutliches Abflachen im Verlauf der Kurven für die maskierten Chromosomen zu beobachten, aber auch eine deutliche Abnahme der Varianz in den Kurvenscharen. Damit sind die Kurven stärker gebündelt und zeigen eine höhere Synchronisation, also ein deutlicheres Korrelationssignal. Dieses Ergebnis ist in gewisser Hinsicht überraschend. Das Maskieren von Sequenzabschnitten mit einer teilweise deutlichen Struktur, also von homogenen Eigenschaften innerhalb einer Klasse von Repeats, führt statt zu einer Erhöhung der Streuung zu einer stärkeren Synchronisation. Die Signatur ändert sich auch in der oben beschriebenen Weise, indem gelegentlich Maxima auf Minima fallen. Auf diese Weise ergeben sich für die Symbolabstände 7 und 21 Minima in den Kurven zu den maskierten Chromosomen, wo sie in der Schar der Korrelationskurven für die nicht maskierten Chromosomen Maxima aufweisen. Umgekehrt stellt die Korrelationsstärke für den Symbolabstand 14 einen Peak für die maskierten Chromosomen dar, während die nicht maskierten Chromosomen ein Tal in der Korrelationskurvenschar in diesem Abstand aufweisen. Die Signatur für beide Spezies ist wie auch bei Ratte und Maus nach der Maskierung verändert. Der Anteil der maskierten Sequenz beträgt für Mensch und Schimpanse jeweils ca. 47%.

Die erste Frage, die es zu beantworten gilt, ist, in welcher Weise die Effekte der Maskierung mit dem prozentualen Anteil von Repeats innerhalb eines Genoms in Verbindung stehen. Es ist ganz klar, dass die reine Menge von repetitiven Elementen einen Einfluss auf die Signatur hat. Es ist aber auch so, dass die Struktur dieser Elemente eine wichtige Rolle spielt. Darauf wird im nächsten Abschnitt ausführlich eingegangen. Die zweite Frage lautet, wie sich die Maskierung auf die Anordnung der Chromosomen im Clusterbaum auswirkt. Abbildung 2.25 zeigt das Ergebnis der Clusteranalyse auf Basis der in Abbildung 2.24 abgebildeten Korrelationskurven der maskierten Chromosomen. Die Aufteilung der Chromosomen in Speziescluster bleibt auch nach der Maskierung in fast allen Fällen erhalten. Eine Veränderung gegenüber dem Clusterbaum in Abbildung 2.19 zeigt sich diesbezüglich nur bei den Chromosomen der Maus und der Ratte. Die Chromosomen dieser Spezies sind nun leicht vermischt, während sie bei unveränderten Chromosomen nahezu vollständig getrennt werden. Dabei haben die Knoten im Subcluster der Chromosomen von Maus und Ratte niedrige Bootstrap-Werte, was eine instabile Substruktur belegt. Diese Eigenschaft bleibt auch erhalten, wenn der betrachtete Bereich von Symbolabständen auf $k = 1, \ldots, 100$ erhöht wird. Die wichtigste Auswirkung der Maskierung repetitiver Elemente auf den Clusterbaum zeigt sich jedoch an noch anderer Stelle: Das Cluster der Chromosomen des Huhns weist eine neue Position im Baum auf, vor der Abzweigung von Mensch, Maus und Ratte und nach der von Drosophila und Moskito. Der Bootstrap-Wert von 99 an diesem Knoten im Baum macht die Robustheit dieser Struktur deutlich. Damit werden, im Gegensatz zu dem auf den originalen Chro-

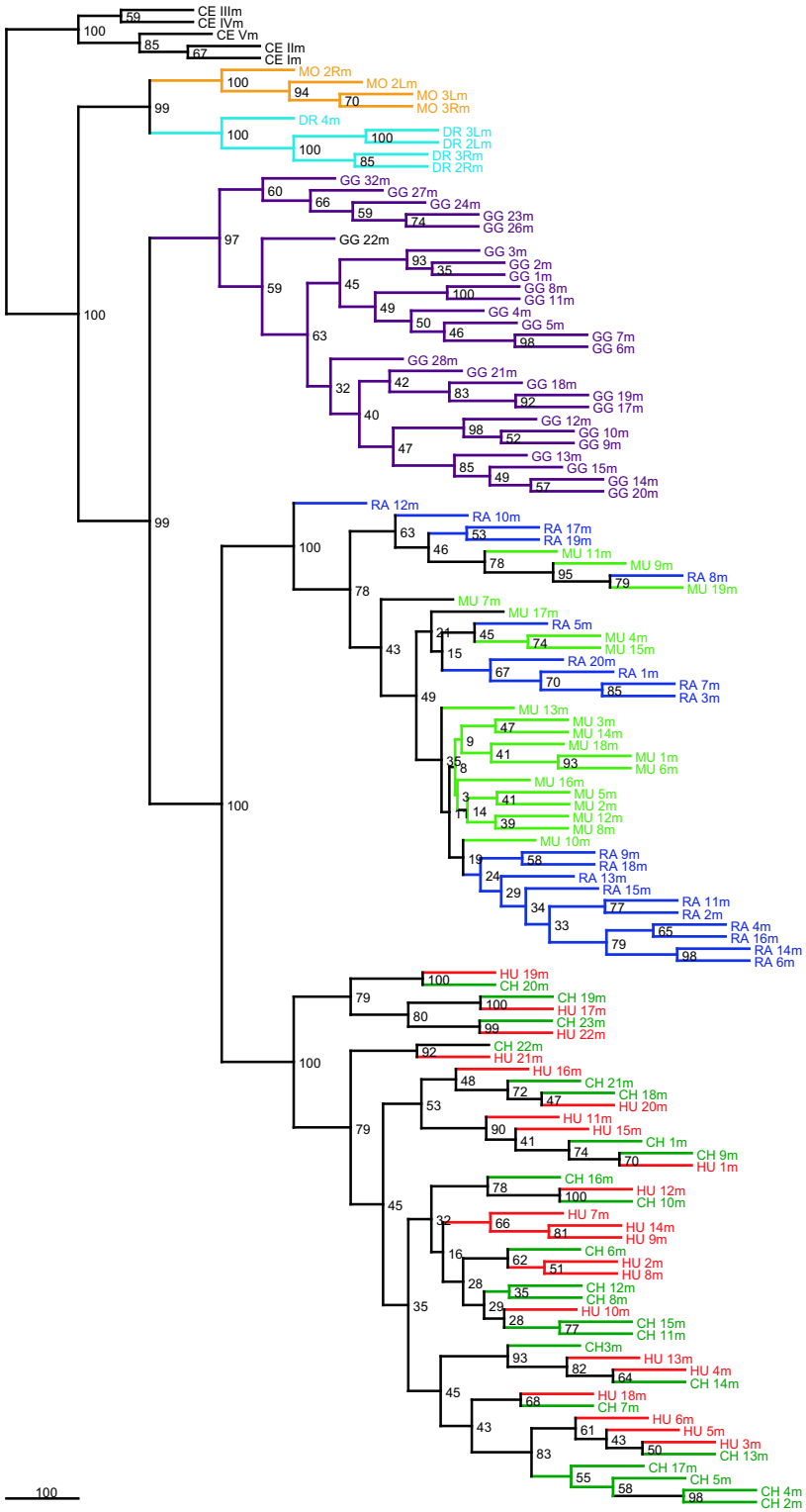

Abb. 2.25. Clusterbaum für Chromosomen der 8 eukaryotischen Spezies aus Abbildung 2.24, deren repetitive Elemente mit zufälligen Symbolsequenzen überschrieben sind (maskiert). Die Abbildung zeigt den *Consensus*-Baum von 100 Bootstrap-Samples mit einer Angabe der Bootstrap-Werte an den Knoten der Zweige.

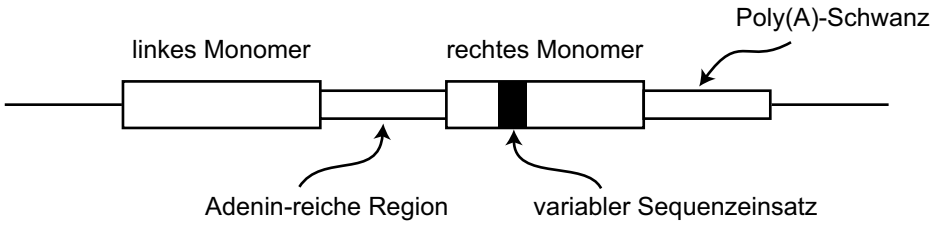

Abb. 2.26. Aufbau eines *Alu*-Repeats.

mosomen basierenden Baum, die Chromosomen des Huhns nun aus phylogenetischer Sicht richtig eingeordnet. Die Vermischung der Chromosomen des Menschen und des Schimpansen bleibt bestehen. Auch treten wieder Paarbildungen aus Chromosomen beider Spezies auf, mit zum Teil ähnlich hohen Bootstrap-Werten. Die beobachteten Paarbildungen entsprechen in vielen Fällen homologen Chromosomen in Mensch und Schimpanse (siehe Tabelle C.5 für eine Zuordnung von homologen Chromosomen). Die Cluster von *C. elegans*, Drosophila und Moskito sind wie auch vor der Maskierung klar getrennt und weisen die gleiche Metastruktur im Baum auf.

2.5 Detailuntersuchung bei Mensch, Maus und Ratte

Für drei Spezies, bei denen eine Maskierung aller annotierten Repeats eine erhebliche Auswirkung auf die Korrelationskurven hat, soll in diesem Kapitel der Frage nachgegangen werden, wieviel Einfluss bestimmte Klassen repetitiver Elemente auf die Korrelationsstruktur haben. Um diesen Sachverhalt zu untersuchen, werden im Folgenden einzelne Klassen von Elementen bei Mensch, Maus und Ratte systematisch eliminiert und der Einfluss auf die Korrelationskurven diskutiert. Die Maskierung erfolgt dabei im Weiteren immer durch Ausschneiden der unterschiedlichen Elemente aus der Sequenz.

2.5.1 Repetitive Elemente: *short interspersed elements*

Die erste Kategorie von mobilen Elementen, die hier betrachtet werden soll, sind SINEs, die je nach Spezies über unterschiedliche Repeat-Familien verfügen. Die SINEs im Genom des Menschen lassen sich in drei Klassen von Elementen unterteilen. Die am besten untersuchte Klasse bildet die primatenspezifische Familie der *Alu*-Repeats. *Alu*-Repeats sind ca. 300 bp lang und finden sich gewöhnlich in Introns, im 3' UTR-Bereich und in intergenischen Regionen (Batzer und Deininger, 2002). Sie besitzen eine zweigeteilte Struktur bestehend aus einem linken und einem rechten Monomer, die durch eine Adenin-reiche Region verbunden sind. Das rechte Monomer weist einen zusätzlichen variablen Sequenzeinsatz von meist 31 bp auf. Ein *Alu*-Element wird außerdem von kurzen Sequenzwiederholungen flankiert, die von den Einsetzstellen stammen. Am 3' Ende findet sich fast immer ein Poly(A)-Schwanz. Abbildung 2.26 zeigt den Aufbau eines solchen Repeats schematisch. *Alu*-Repeats stammen aus 7SL RNA und können sich nicht selbstständig vermehren, sondern nutzen zur Retrotransposition die Maschinerie der LINEs, einer weiteren

Klasse von repetitiven Elementen (Dewannieux et al., 2003). Sie führen unter anderem zu Mutationen, Rekombinationen, *gene conversion* und alternativem Spleißen (siehe Batzer und Deininger (2002) für einen ausführlichen Überblick). Ca. 10% des menschlichen Genoms besteht aus *Alu*s (Human Genome Sequencing Consortium, 2001). Neben den *Alu*s wird außerdem die Klasse der *mammalian-wide interspersed repeats* (MIRs) den SINEs des Menschen zugeordnet. Die MIRs werden in die Klassen MIR und MIR3 eingeteilt und kommen auch in anderen Säugetieren vor. Diese Repeats haben eine Länge von mindestens 260 Basen (Murnane und Morales, 1995). Ihre Erkennung ist schwierig, bedingt durch das hohe Alter und die damit verbundene Divergenz der Sequenzen. 3 % des menschlichen Genoms wird den MIR/MIR3 Familien zugeordnet (Human Genome Sequencing Consortium, 2001).

Der prozentuale Anteil von SINEs bei der Maus beträgt ca. 8% (Mouse Genome Sequencing Consortium, 2002). Die als B1-Familie bezeichnete Klasse von SINEs entspricht der Familie der *Alu*s im menschlichen Genom und hat einen Anteil von 2.66%. Die MIR und MIR3-Familien stellen zusammen 0.57% der Sequenz dar. Neben diesen Repeats, die auch im menschlichen Genom vorkommen, gibt es bei der Maus die Familien B2 (2.39%), B4 (2.36%) und ID (0.25%), die keine äquivalente Entsprechung im menschlichen Genom haben. Der prozentuale Anteil von SINEs bei der Ratte beträgt ca. 7% (Rat Genome Sequencing Project Consortium, 2004), wobei die Familien von B1 (1.65%), B2 (2.15%), B4 (2.15%), ID (0.76%) und MIR (0.51%) sich damit nur gering in ihrem Anteil von dem in der Maus unterscheiden. Schon aus diesen sehr unterschiedlichen Häufigkeiten und internen Homologien der SINEs ist zu erwarten, dass die Maskierung bei den Chromosomen des Menschen eine größere Wirkung auf die Korrelationskurven haben wird. Abbildung 2.27 zeigt die Korrelationskurven der SINEs-maskierten Autosomen des Menschen, der Maus und der Ratte und als Referenzen die jeweiligen Korrelationskurven, die man für die unmaskierten Chromosomen erhält.

In Abbildung 2.27 a sind die Korrelationskurven für die SINEs-maskierten Chromosomen des Menschen aufgetragen. Die Korrelationskurven der maskierten Chromosomen sind denen der unmaskierten ähnlich, zeigen aber einen anderen Verlauf als die der Originalsequenzen. Als Erstes fällt auf, dass die Korrelationskurven nach der Maskierung stärker gebündelt sind, also eine geringere Varianz aufweisen. Die Synchronisation innerhalb der maskierten Chromosomen in Bezug auf diese hier diskutierten Symbolkorrelationen ist also größer. Als Zweites sieht man, dass zwar beide Kurvenscharen − bis auf wenige Ausnahmen − die gleiche Abfolge von Höhen und Tiefen aufweisen, aber die Amplitude der maskierten Sequenzen geringer ist, und die Kurvenschar − besonders in den Abständen von 15-30 Basen − flacher als die der unmaskierten Chromosomen wirkt. Dadurch kommt es nur zu einer geringen Überlagerung der zwei Arten von Korrelationskurven, die mit größerem Abstand immer mehr abnimmt. Die Korrelationsstärke im Symbolabstand $k = 1$ unterscheidet sich − bedingt durch die hohe Varianz − nicht signifikant für die maskierten Chromosomen und unmaskierten Chromosomen. Die Tatsache, dass die Löschung systematisch strukturierter Sequenzabschnitte aus den Chromosomen zu einer stärkeren Synchronisation der Korrelationskurven führt, stellt eine Überraschung dar.

Abbildung 2.27 b zeigt die Korrelationskurven für die SINE-maskierten und unmaskierten Chromosomen der Maus. Der Unterschied zwischen den beiden Kurvenscharen ist weniger ausgeprägt als beim Menschen. Sie zeigen die gleiche Abfolge von Peaks und liegen in vielen Symbolabständen übereinander. Aber auch hier stellt man eine verminderte Varianz der Korrelationskurven fest, die auf eine höhere Synchronisation der maskierten Chromosomen zurückzuführen ist.

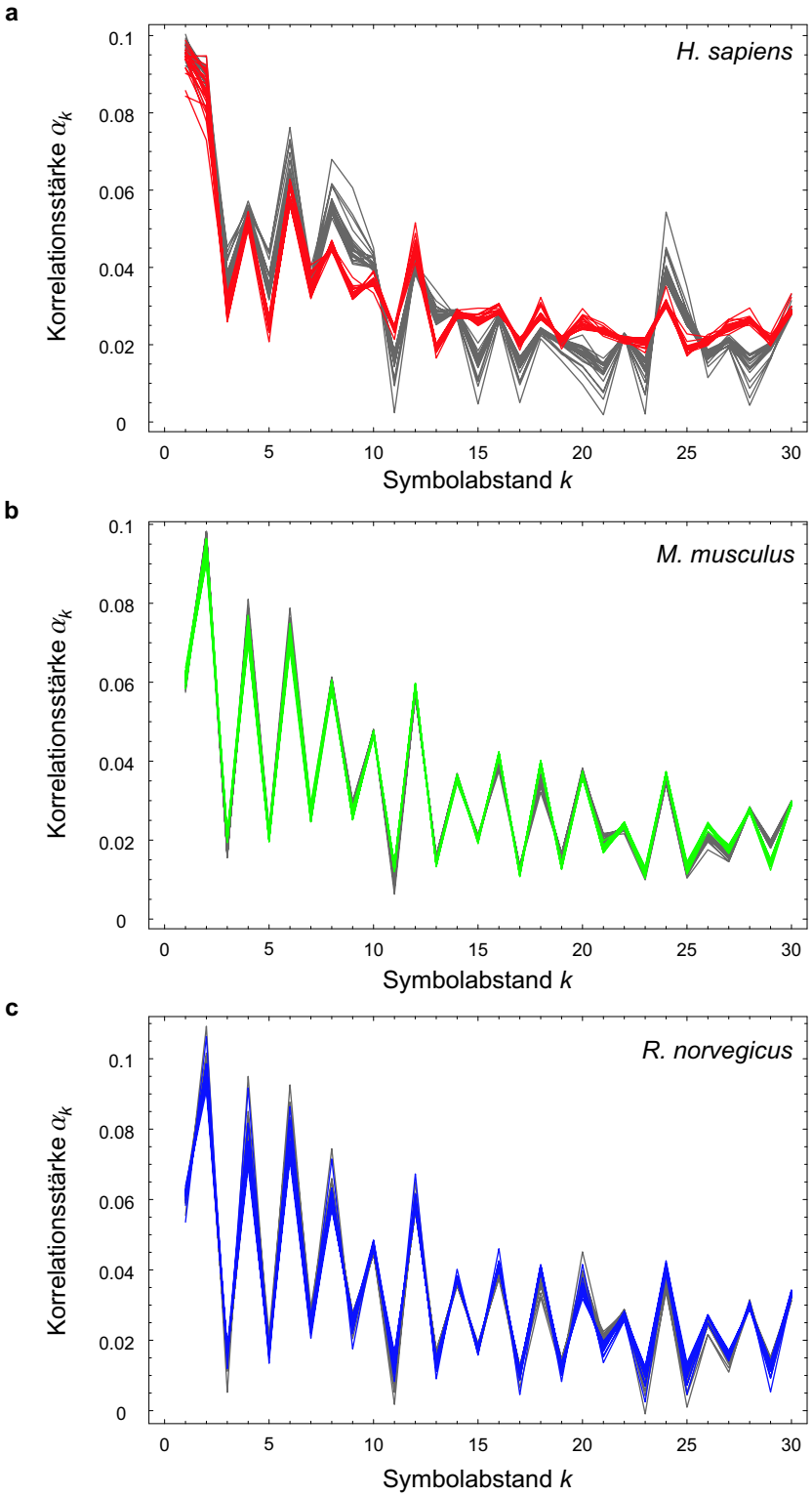

Abb. 2.27. Korrelationskurven für die Chromosomen von **a** *H. sapiens*, **b** *M. musculus* und **c** *R. norvegicus* nach der Maskierung der zu der Klasse der SINEs gehörenden repetitiven Elemente [farbig], im Vergleich zu den Korrelationskurven, die für die unmaskierten Sequenzen erhalten werden [grau].

Im Fall der Ratte, deren Korrelationskurven für die maskierten und unmaskierten Chromosomen in Abbildung 2.27 c dargestellt sind, ist ein ähnliches Verhalten wie bei der Maus zu beobachten. Die Korrelationskurven sind für unmaskierte und maskierte Chromosomen sehr ähnlich. Letztere weisen eine verminderte Varianz auf, also eine leicht bessere Synchronisation der Kurven.

Eine spezieseinheitliche Signatur der Chromosomen bleibt auch nach der Maskierung für alle drei betrachteten Spezies erhalten. Es ist anzunehmen, dass der unterschiedliche prozentuale Anteil der SINEs im Genom des Menschen, der Maus und der Ratte sowie die höhere interne Ähnlichkeit der SINEs beim Menschen die Hauptgründe für die unterschiedlich starke Auswirkung der Maskierung auf die jeweiligen Korrelationskurven darstellen. Der Unterschied der prozentualen Anteile von Retrotransposons im Menschen, der Maus und der Ratte reflektiert auch die höhere Divergenz in Nagetieren im Vergleich zum Menschen, welches eine Identifikation älterer Transposons unmöglich macht (Mouse Genome Sequencing Consortium, 2002; Deininger et al., 2003).

2.5.2 Repetitive Elemente: *long interspersed elements*

LINEs gehören zu der Gruppe der Retrotransposons, die den Hauptteil der Transposons in Säugetieren darstellen. Retrotransposons mobilisieren sich durch die Codierung einer Endonuklease und einer reversen Transkriptase (Deininger et al., 2003) und stellen eine autonome Einheit von Elementen dar. Die LINEs im Genom des Menschen werden in die Klassen *L*1, *L*2 und *L*3 unterteilt, die zusammen ca. 20% des Genoms ausmachen (Human Genome Sequencing Consortium, 2001). Die größte Subklasse stellen die *L*1-Elemente (16.89%), gefolgt von den *L*2- (3.22%) und L3-Elementen (0.31%). Die Maskierung der LINEs im menschlichen Genom führt auf die in Abbildung 2.28 a dargestellten Korrelationskurven. Im Vergleich zu den unmaskierten Chromosomen zeigen diese eine leicht stärkere Amplitude für die einzelnen Abstände. Damit ist − im Gegensatz zu den Beobachtungen bei den *Alu*-Repeats − diesmal eine Verstärkung der Korrelationsstruktur zu beobachten. Die Varianz der Kurven bleibt aber gleich oder nimmt sogar leicht zu, und somit tritt keine stärkere Synchronisation der Korrelationskurven ein.

Auch bei den in Abbildung 2.28 b und c dargestellen Korrelationskurven für die LINE-maskierten Chromosomen der Maus und der Ratte lässt sich eine Verstärkung der Abfolge von Höhen und Tiefen feststellen, die jedoch ähnlich moderat ausfällt wie beim Menschen. Die Scharen der Korrelationskurven zeigen eine ähnlich große Varianz und somit eine ähnliche Synchronisation.

Neben den LINEs zählen auch die LTRs zu den Retrotransposons. Im Genom des Menschen, der Maus und der Ratte werden 4 Klassen unterschieden, die insgesamt 8.29%, 9.87% bzw. 9.04% des Genoms ausmachen. Eine Maskierung dieser Elemente führt zu keiner signifikanten Änderung der Korrelationskurven. Der Grund liegt vermutlich in der geringen Homologie der Elemente. Außerdem gibt es in allen drei Spezies noch DNA-Transposons, die hier nicht als separate Klasse untersucht wurden.

2.5.3 Repetitive Elemente: Mikrosatelliten

Mikrosatelliten werden auch als *short tandem repeats* (STRs) oder *simple sequence repeats* (SSRs) bezeichnet. Mikrosatelliten gehören zu den variabelsten Arten von DNA im Genom, die ihre Unterschiedlichkeit hauptsächlich aus der Variation ihrer Länge beziehen. Ein Mikrosatellit wird im

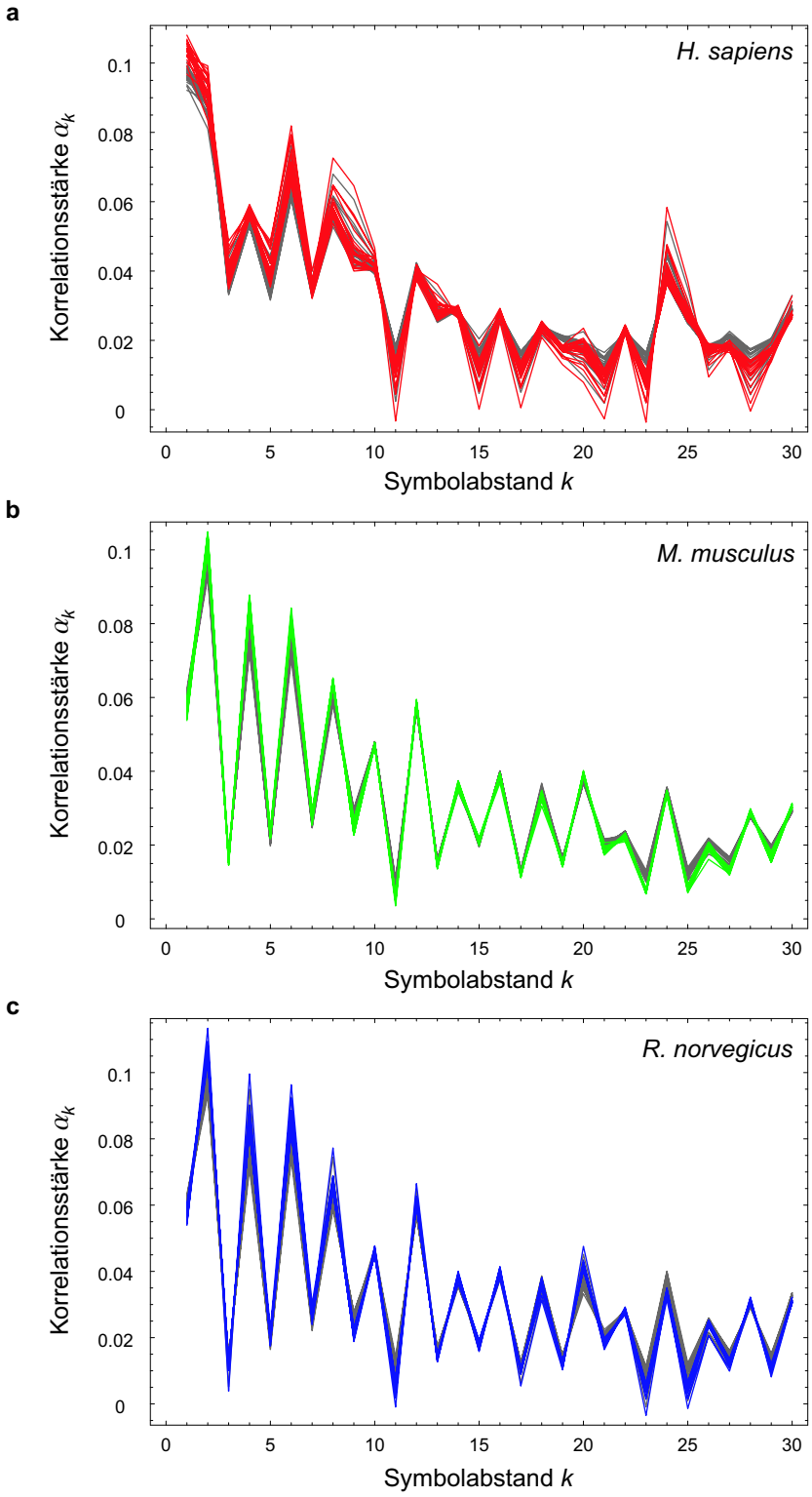

Abb. 2.28. Korrelationskurven für die Chromosomen von **a** *H. sapiens*, **b** *M. musculus* und **c** *R. norvegicus* nach der Maskierung der zu der Klasse der LINEs gehörenden repetitiven Elemente [farbig], im Vergleich zu den Korrelationskurven, die für die unmaskierten Sequenzen erhalten werden [grau].

Allgemeinen durch ein Motiv, seine Länge und die Anzahl seiner Wiederholungen charakterisiert. Mono-, Di-, Tri- und Tetra-Nukleotid Repeats stellen die Haupttypen von Motiven dar, aber auch Wiederholungen von fünf (Penta-) oder sechs (Hexa-) Nukleotiden werden als Mikrosatelliten klassifiziert (Ellegren, 2004). Man bezeichnet Wiederholungen von längeren Motiven häufig als Minisatelliten, und im Extremfall als Satelliten-DNA. Die Annotation der Mikrosatelliten in der Datenbank des *Genome Browser* basiert auf Resultaten der Software *Repeatmasker* (Smit et al., 2004). Ein Anwachsen oder Abbau von Mikrosatelliten wird in der Regel mit Fehlern bei der Replikation von DNA in Verbindung gebracht (Kunkel und Bebenek, 2000; Ellegren, 2004). Dabei kommt es zur Disassoziation der DNA-Polymerase vom Template-Strang und einer falschen Wiederanlagerung, was zur Einfügung oder Löschung von Nukleotiden relativ zum Template-Strang führt. In codierender DNA werden die meisten dieser Fehler durch Reparaturmechanismen behoben und nur wenige führen zu einer Mutation. Der größte Teil an Mikrosatelliten liegt in nichtcodierenden Bereichen, entweder in intergenischen Bereichen oder Introns. In codierender DNA können Mikrosatelliten zum Verlust von Genfunktion führen (Li et al., 2004).

Es wäre zu erwarten, dass die Menge und statistische Prägung von Mikrosatelliten in Säugetieren aufgrund der Erwartung ähnlicher elementarer Prozesse vergleichbar ist, was aber überraschenderweise nicht zutrifft (Beckman, 1992; Mouse Genome Sequencing Consortium, 2002). Die Menge an Simple-Repeats beträgt im Genom des Menschen 0.87%, im Genom der Maus 2.41% und im Genom der Ratte 2.38%. Damit erhält man bei den gleichen Einstellungen der Analysewerkzeuge für das Genom der Maus und der Ratte zwei- bis dreimal so viele Mikrosatelliten wie für das menschliche Genom. Die Mikrosatelliten bei Maus und Ratte sind im Besonderen länger als beim Menschen (Mouse Genome Sequencing Consortium, 2002; Rat Genome Sequencing Project Consortium, 2004).

Die Maskierung aller Mikrosatelliten im menschlichen Genom führt auf eine Änderung der Korrelationskurven, wie in Abbildung 2.29 a zu sehen ist. Die Korrelationskurven der maskierten Chromosomen zeigen einen leicht flacheren Verlauf als die der Originalchromosomen. Die Varianz der maskierten Kurvenschar bleibt jedoch ungefähr gleich und beide Kurvenscharen zeigen die gleiche Signatur. Die Abweichungen sind über den ganzen Vektor verteilt, wobei kleinere Symbolabstände die größten Abweichungen zeigen. Bedenkt man jedoch, dass die Mikrosatelliten im menschlichen Genom nur 0.87% der Masse ausmachen, so ist der Einfluss der Maskierung bemerkenswert.

In Abbildung 2.29 b sind die Korrelationskurven für die Simple-Sequence-Repeats-maskierten Chromosomen der Maus zusammen mit den Originalkurven aufgetragen. Die maskierten Chromosomen zeigen einen deutlich flacheren Verlauf in den Korrelationskurven. Klare Abweichungen treten über alle Symbolabstände auf. Die Streuung innerhalb der Kurvenscharen ist in beiden Fällen gering und unterscheidet sich kaum. Für die Ratte ergibt sich ein sehr ähnliches Bild in Abbildung 2.29 c.

Der Anteil von Mikrosatelliten beträgt bei der Maus 2.41% und bei der Ratte 2.38%. Die Auswirkungen auf die Korrelationskurven sind in beiden Fällen beträchtlich im Verhältnis zur maskierten Sequenzmenge. Damit wird deutlich, dass Mikrosatelliten bei diesen beiden Spezies einen sehr großen Beitrag zu der hier gemessenen Korrelationsstruktur leisten.

Es zeigt sich mit diesem Ergebnis, dass die Korrelationsstruktur bei Säugetieren zu einem großen Teil von repetitiver DNA hervorgerufen wird. Dabei ist zu beobachten, dass im Genom des Menschen die Signatur in besonderem Maße durch SINEs beeinflusst wird. Für Maus und Ratte ergibt

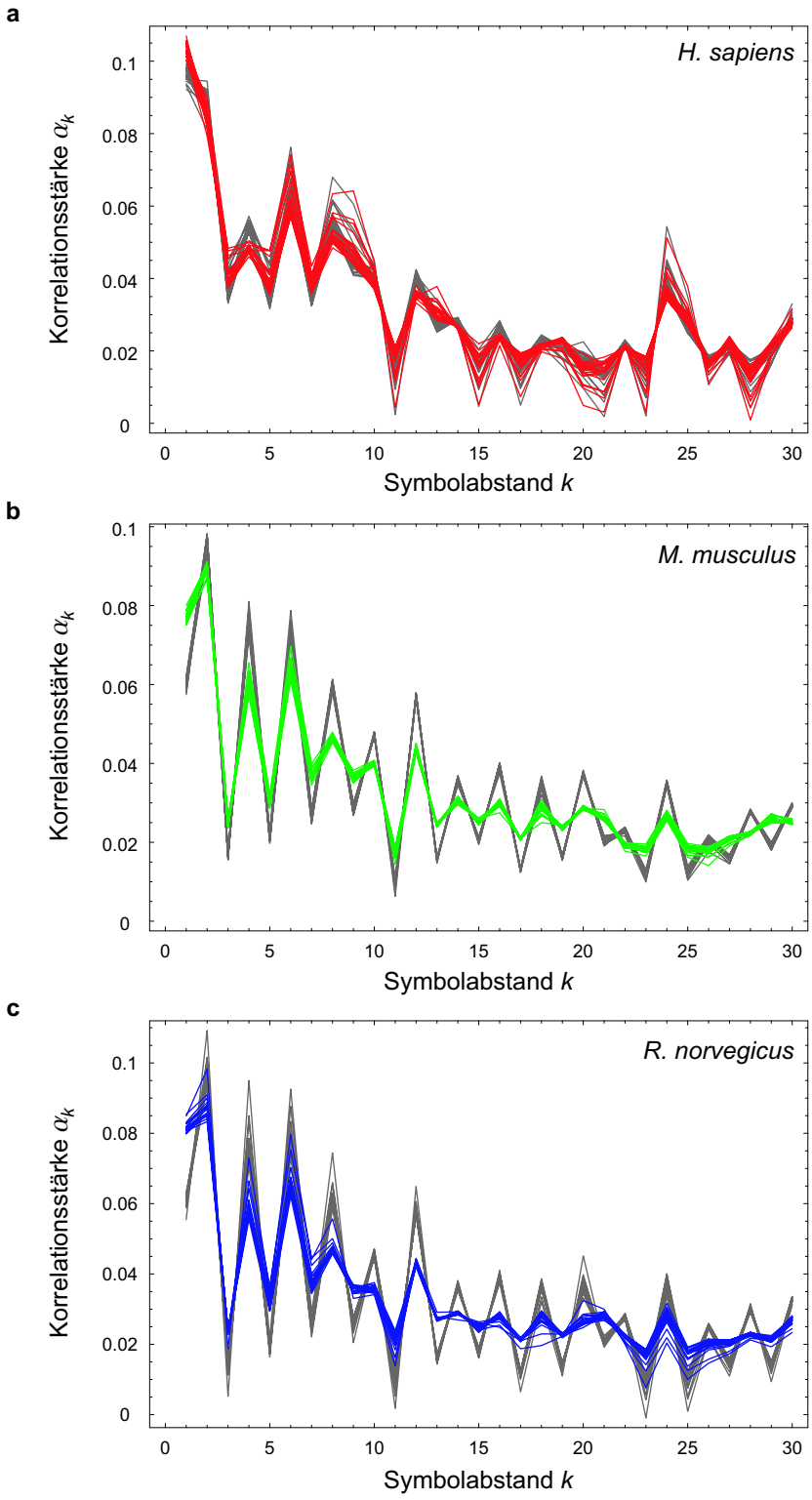

Abb. 2.29. Korrelationskurven für die Chromosomen von **a** *H. sapiens*, **b** *M. musculus* und **c** *R. norvegicus* nach der Maskierung der Mikrosatelliten [farbig], im Vergleich zu den Korrelationskurven, die für die unmaskierten Sequenzen erhalten werden [grau].

sich ein großer Beitrag zur Korrelationsstruktur aus den Mikrosatelliten. Es ist bekannt, dass die Verteilung von Mikrosatelliten sich innerhalb der Eukaryoten unterscheidet. Nagetiere weisen eine höhere Dichte von Mikrosatelliten auf, diese sind länger als beim Menschen und haben eine höhere interne Ähnlichkeit (Mouse Genome Sequencing Consortium, 2002; Rat Genome Sequencing Project Consortium, 2004; Almeida und Penha-Goncalves, 2004). Es gibt jedoch bisher keine Erklärung für diesen Sachverhalt. Außerdem wird beobachtet, dass auch nah verwandte Spezies Unterschiede in der Verteilung von Mikrosatelliten aufweisen (Webster et al., 2002). LINEs haben dagegen in beiden Spezies nur einen geringen Einfluss auf die Kurven. Dies liegt sicher auch an der geringen homogenen Struktur innerhalb der LINEs für Mensch, Maus und Ratte. Für Korrelationen im Genom des Menschen, der Maus und der Ratte kann somit eine quantitative Verbindung zwischen statistischen Eigenschaften und biologischen Kenngrößen der Sequenz hergestellt werden.

Das Verbleiben von ordnenden Prinzipien in der Korrelationsstruktur nach Löschung aller repetitiver Elemente durch Ausschneiden aus der Sequenz zeigt, dass nicht ausschließlich die Ebene der Homogenität von repetitiven Elementen und ihrer Verteilung durch die Korrelationskurven gemessen werden.

3

Schlussfolgerungen und Ausblick

In aller Konsequenz wird der Schritt hin zur rein statistischen Analyse einer DNA-Sequenz durch Anwendung von Methoden der Informationstheorie vollzogen. Dort werden statistische Korrelationen zwischen Symbolen der DNA-Sequenz sichtbar gemacht und mit biologischer Funktion in Verbindung gebracht. Auf diese Weise konnte in der hier vorliegenden Arbeit eine neue, auf Symbolkorrelationen basierende Genom-Signatur formuliert werden. Um dieses Phänomen speziesabhängiger Korrelationen sichtbar machen zu können, wird eine besondere mathematische Methode zur Quantifizierung solcher Symbolkorrelationen angewendet. Die Kernidee dabei ist, einen Markov-Prozess höherer Ordnung durch ein statistisches Schätzverfahren an die Symbolsequenz anzupassen und die Prozessparameter als Maß für die Korrelationsstärke zu verwenden. Die so beobachtbaren Korrelationen bzw. Korrelationskurven können mit Hilfe von Clusteranalysen untersucht werden. Dabei stellt jedes Chromosom einer Spezies ein eigenes Taxon dar. Die Anordnung der sich ergebenden Chromosomencluster spiegelt teilweise phylogenetische Eigenschaften der untersuchten Spezies wider. Im Falle der Spezies *C. elegans*, *D. melanogaster*, *A. gambiae*, *H. sapiens*, *M. musculus* und *R. norvegicus* ergeben sich fast vollständig getrennte Cluster von Chromosomen einer Spezies, die sich interpretierbar im Baum anordnen. Dabei zeigt sich insbesondere eine klare Trennung zwischen *C. elegans*, Insekten und Säugetieren. *P. troglodytes* als naher Verwandter des Menschen tritt bei einer um diese Spezies erweiterten Analyse mit seinen Chromosomen nicht als separates Cluster im Baum auf, sondern in einem gemeinsamen Cluster mit den Chromosomen des Menschen. Die dabei häufig auftretenden robusten Paarungen von Chromosomen beider Spezies sind in enger Übereinstimmung zu den aus biologischer Sicht orthologen Chromosomen. Bei einer Nukleotid-Divergenz beider Spezies von ca. 1% auf der Ebene des gesamten Genoms (The Chimpanzee Sequencing and Analysis Consortium, 2005) ist es nicht überraschend, dass eine Trennung dieser Spezies auf Basis der Korrelationsstruktur nicht möglich ist.

Bei Erweiterung der Analyse um *G. gallus* bilden die Chromosomen dieser Spezies ein separates Cluster, das zusammen mit den anderen in der Analyse untersuchten Wirbeltieren eine Substruktur im Baum ergibt. Die Position des Huhns innerhalb dieser Struktur liegt neben den Chromosomen des Menschen (und des Schimpansen) und nicht, wie aus phylogenetischer Sicht zu erwarten wäre, vor dem Cluster der Säugetiere. Die Schar der Korrelationskurven des Huhns ist der des Menschen ähnlich, zeigt aber trotzdem deutliche Unterschiede.

Ein in Bezug auf bestehende Forschungsdebatten um die Längenskalen, auf denen Speziesinformationen vorliegen, wichtiges Resultat stellt die Zunahme an Speziestrennung mit wachsendem Korrelationsbereich dar. Dies wurde hier am Beispiel von Maus und Ratte nachgewiesen.

Um zu untersuchen, welche funktionell benennbaren Bestandteile der DNA-Sequenz für die beobachtete Intraspezies-Synchronisation und die Interspezies-Unterschiede verantwortlich sein könnten, werden unterschiedliche biologisch abgrenzbare Bereiche der DNA-Sequenz eliminiert. Eine solche Maskierung von Teilen der Sequenz wird durch das Ausschneiden oder Überschreiben dieser Abschnitte mit zufälligen Symbolsequenzen realisiert. Für ein Verständnis der Korrelationsstruktur hat sich die Maskierung repetitiver DNA als besonders aufschlussreich erwiesen. Bedingt durch den hohen prozentualen Anteil von repetitiver DNA in Eukaryoten (im Menschen bis zu 50% des Genoms (Human Genome Sequencing Consortium, 2001)), konnte ein Einfluss dieser Elemente auf die Korrelationsstruktur erwartet werden. Im ersten Schritt wurden alle bekannten repetitiven Elemente in den Spezies maskiert. Dabei zeigte sich ein teilweise großer Einfluss auf die Korrelationsstruktur der Chromosomen. Für die Spezies *C. elegans*, *D. melanogaster*, *A. gambiae*, und *G. gallus* beläuft sich der maskierte Anteil an der Gesamtmenge an sequenzierter DNA auf 9-17%. Diese Korrelationskurven werden nur geringfügig beeinflusst. Für die in der Analyse untersuchten Säugetiere hat die Maskierung einen deutlichen Einfluss, indem die Amplitude der Korrelationskurven stark sinkt, und im Fall von Mensch und Schimpanse eine deutlich stärkere Synchronisation der Kurven eintritt. Die Chromosomen der Maus und der Ratte ausgenommen, führt eine Clusteranalyse nach der Maskierung zu einer klaren Trennung der Spezies im Baum. Dies ist ein Beleg dafür, dass die Information in den Korrelationskurven nur partiell von repetitiver DNA getragen wird. Ein „Bereinigen“ der evolutionären Einflüsse von repetitiven Elementen führt auf eine erhebliche Änderung der Position des Huhns im Clusterbaum. Die neue Position des Huhns steht im Einklang mit phylogenetischen Erwartungen.

Die Ergebnisse der Maskierung legen die Vermutung nahe, dass die reine Menge an repetitiver DNA im Genom einen großen Einfluss auf die Resultate hat. Deshalb wird am Beispiel des Menschen, der Maus und der Ratte eine differenzierte Vorgehensweise bei der Maskierung angewendet. Durch das systematische Eliminieren einzelner Klassen von repetitiven Elementen kann so gezeigt werden, dass für unterschiedliche Spezies unterschiedliche Klassen den jeweils größten Beitrag zur Korrelationsstruktur leisten. Im Genom des Menschen zeigt das Maskieren von *Alu*-Repeats den deutlichsten Einfluss auf die Signatur, für Maus und Ratte sind es die Mikrosatelliten. Die Signatur wird damit zu einem großen Teil durch die interne Ähnlichkeit der repetitiven Elemente, also ihre Homogenität, und durch ihre Häufigkeitsverteilung bestimmt, und weniger durch ihre reine Menge. Im Kern liefern die Untersuchungen zur Maskierung einzelner Klassen repetitiver Elemente zwei überraschende Befunde: Zum einen gibt es Fälle, bei denen eine Maskierung die Systematik der Korrelationskurven erhöht. Hierzu gehört die Verminderung der Varianz in den entsprechenden Kurvenscharen und die nach einer Maskierung phylogenetisch plausible Einordnung der Korrelationsstruktur des Huhns. Zum anderen beobachtet man eine Verminderung der Systematik durch die Maskierung. Ein Beispiel dafür ist die Reduktion von Speziestrennbarkeit nach Maskieren der repetitiven Elemente von Maus und Ratte.

Diese Befunde stellen die methodische und inhaltliche Grundlage für eine modellhafte, dynamische Betrachtung der Genom-Evolution dar, bei der Prozesse wie segmentielle Duplikation, Mutation, die Dynamik von Mikrosatelliten und Retrotransposition iteriert werden und in ihren Auswirkungen auf die Korrelationsstruktur eines solchen „simulierten Genoms“ quantitativ analysiert werden können. Die Korrelationsstruktur eines Chromosoms wird so als eine evolutionäre

Prozesssignatur betrachtet. Unterschiede zwischen Spezies lassen sich in einer solchen, modellierenden Betrachtung auf Unterschiede zwischen den Prozessen zurückführen. Methodisch legt diese Arbeit die Grundlage für diesen grundlegenden Weg einer neuen, an Simulationen und dynamischen Modellen orientierten Systembiologie, indem eine verfeinerte Beschreibungsform statistischer Symbolkorrelationen entwickelt wurde. Auf der inhaltlichen Ebene konnte zum einen gezeigt werden, dass eine ausreichende Zahl systematischer Speziesunterschiede vorhanden ist, um der Hypothese über prozessuale Unterschiede nachgehen zu können. Zum anderen wurde hier aber auch die Prägung der Korrelationsstruktur durch repetitive Elemente direkt nachgewiesen und gezeigt, dass solche Korrelationen über große Abstände Speziesinformation tragen.

4

Zusammenfassung

Auf einer genomweiten Skala besitzen eukaryotische DNA-Sequenzen eine mosaikhafte Struktur, eine komplexe Abfolge aus Genen, nicht-codierenden Wiederholungen von Genen (Pseudo-Genen) und repetitiven Sequenzen, die durch scheinbar zufällige Segmente verbunden sind. Die unterschiedliche Struktur dieser Elemente führt zu Symbolkorrelationen, also statistischen „Abhängigkeiten“ zwischen den Basen. In der vorliegenden Arbeit konnte gezeigt werden, dass solche Korrelationen ein unerwartet starkes, innerhalb der Chromosomen einer Spezies hoch synchronisiertes Signal darstellen. Alle Chromosomen einer Spezies zeigen das gleiche charakteristische Muster, welches sich signifikant von denen anderer Spezies unterscheidet. Dabei konnte nachgewiesen werden, dass dieses Korrelationsmuster nicht ausschließlich von Dinukleotiden induziert wird, da sich bei der Betrachtung größerer Symbolabstände eine deutliche Zunahme der artspezifischen Information ergibt. Die gemessene Korrelationsstruktur weist auf der Ebene ganzer Chromosomen über die hohe Synchronisation innerhalb einer Spezies hinaus außerdem ein weiteres Ordnungsprinzip auf: Auf Basis der kurzreichweitigen Korrelationen gewonnene Clusterbäume zeigen eine Übereinstimmung mit der Phylogenie der beteiligten Spezies. In eukaryotischen Genomen wird ein großer Teil der DNA repetitiven Elementen zugeordnet. Die Maskierung dieser Elemente als mögliche Träger von speziesspezifischer Information führt zu einer Änderung der beobachteten Genom-Signaturen. Dennoch bleibt nach der Maskierung aller repetitiven Elemente eine artspezifische Speziessignatur erhalten. Für Korrelationen im Genom von *H. sapiens*, *M. musculus* und *R. norvegicus* konnte zudem eine quantitative Verbindung zwischen diesen statistischen Eigenschaften und biologischen Kenngrößen der Sequenz hergestellt werden. Die systematische Maskierung verschiedener Klassen repetitiver Elemente wirkt sich dabei unterschiedlich auf die Korrelationsstruktur aus. So zeigt sich etwa, dass Mikrosatelliten bei *M. musculus* und *R. norvegicus* den größten Beitrag liefern, während bei *H. sapiens short interspersed elements* (SINEs) die Korrelationsstruktur stark beeinflussen.

A

Mathematische Eigenschaften der DAR(p)-Prozesse

Ein DAR(p)-Prozess kann zur Erzeugung von Symbolsequenzen mit einer Markov-Eigenschaft höherer Ordnung herangezogen werden. Solche Realisierungen eines stochastischen Prozesses können zum Beispiel zum Test oder zur Eichung von Werkzeugen wie der Transinformation oder Entropien höherer Ordnung verwendet werden. Umgekehrt führt die empirische Analyse solcher Sequenzen zu einem besseren Verständnis der zugrunde liegenden Prozesse. Im Folgenden werden informationstheoretische Maße eingesetzt, um die Parameter des DAR(p)-Prozesses besser zu verstehen, sowie auch die informationstheoretischen Maße, mit denen die Realisierungen untersucht werden. Außerdem wird der DAR(p)-Prozess noch einmal formal definiert und es werden einige analytische Überlegungen angestellt.

A.1 Verallgemeinerung der Shannon-Entropie

Eine Verallgemeinerung der Shannon-Entropie aus Gleichung (1.7) sind Entropien höherer Ordnung. Auf Wahrscheinlichkeiten für das Beobachten bestimmter Subsequenzen der Länge n (n-Worte oder n-Blöcke) innerhalb einer Sequenz formuliert man dazu eine Entsprechung zu Gleichung (1.7). Die Wahrscheinlichkeit für das Beobachten einer Subsequenz $x_1,\ldots,x_n$ mit $x_i \in \Sigma$ wird mit $p(x_1,\ldots,x_n)$ bezeichnet. Die Größen

$$H_n = -\sum_{(x_1,\ldots,x_n)\in\Sigma^n} p(x_1,\ldots,x_n)\log_\lambda p(x_1,\ldots,x_n) \tag{A.1}$$

sind dann die n-Block-Entropien oder *Entropien höherer Ordnung* (Ebeling et al., 1998). Die Summe in Gleichung (A.1) läuft über alle möglichen n-Worte. Äquivalent zur Shannon-Entropie beschreiben diese verallgemeinerten Entropien die mittlere Unsicherheit bei der Beobachtung eines n-Wortes bei einer zugrunde liegenden Verteilung P bzw. die benötigte mittlere Information, um ein n-Wort vorherzusagen.

Um die Frage zu beantworten, wieviel Information im Mittel benötigt wird, um von einem bekannten n-Wort ausgehend, ein Wort der Länge $n+1$ vorherzusagen, geht man zu Differenzen benachbarter H_n über. Damit erhält man ein Maß für die Informationsänderung beim Wechsel von n-Worten zu $(n+1)$-Worten. Betrachtet man also statt der H_n Differenzen der Form

$$h_n = H_{n+1} - H_n, \quad h_0 := H_1\ , \tag{A.2}$$

so erhält man genau diesen Informationsgehalt, nämlich die Steigung für jedes n. Diese Größen h_n werden als *bedingte Entropien* bezeichnet (Ebeling et al., 1998).

A.2 DAR(p)-Prozesse

Der DAR(p) Prozess ist wie folgt definiert (Jacobs und Lewis, 1978, 1983):
Sei $A = \{a_1,\ldots,a_\lambda\}$ ein Alphabet mit λ Buchstaben, $\lambda \in \mathbb{N}$, $\lambda \geq 1$. Sei außerdem $\{Y_n\}$ eine Folge von unabhängig und identisch verteilten Zufallsvariablen einer Marginalverteilung π mit Werten in $I \subseteq \mathbb{N}$ für die gilt:

$$P(Y_n = a_i) = \pi(a_i)\ , \quad a_i \in A\ . \tag{A.3}$$

Sei $\{V_n\}$ eine unabhängige Folge Bernoulli-verteilter Zufallsvariablen für die

$$P(V_n = 1) = 1 - P(V_n = 0) = \rho \quad \text{mit} \quad 0 \leq \rho < 1 \tag{A.4}$$

gilt. Sei $\{A_n\}$ eine Folge von unabhängigen Zufallsvariablen mit Werten in $\{1, 2, ..., p\}$, für die gilt

$$P(A_n = i) = \alpha_i \geq 0 \quad \text{mit} \quad i = 1, 2, ..., p \quad \text{und} \quad \sum_{i=1}^{p} \alpha_i = 1 \,. \tag{A.5}$$

Eine Folge $\{X_n\}$, welche durch

$$X_n = V_n X_{n-A_n} + (1 - V_n) Y_n \quad \text{mit} \quad n = p,\, p+1,\, p+2,\, \ldots \tag{A.6}$$

bestimmt ist, heißt DAR(p) Prozess.

Als Erstes untersuchen wir DAR(1)-Prozesse, deren Folge $\{X_n\}$ bestimmt wird durch

$$X_n = V_n X_{n-1} + (1 - V_n) Y_n \,, \qquad n = 1, 2, \ldots \tag{A.7}$$

und verallgemeinern die Ergebnisse dann auf DAR(p)-Prozesse mit $p > 1$. Es kann mit fixierten deterministischen Werten $x_0, \ldots, x_{p-1}$ gestartet werden, oder die Startwerte können als Realisierungen der Zufallsvariablen $X_0, \ldots, X_{p-1}$ nach der Marginalverteilung gezogen werden. Die zweite Methode stellt die Stationarität des gesamten Prozesses sicher, nicht erst für $n \to \infty$, wie bei der ersten. In Abbildung A.1 sieht man die bedingten Entropien h_n für verschiedene $\rho \in [0, 1]$. Die Sequenzlänge L beträgt mit 1.05×10^6 ungefähr λ^n mit $\lambda = 4$ und $n = 10$. Die Länge L der Sequenz ist damit so gewählt, dass sie gerade die Anzahl der verschiedenen n-Worte darstellt, die man (theoretisch) bei überlappender Zählung in der Sequenz finden kann. Man erkennt sofort die charakteristische Signatur einer Markov-Sequenz erster Ordnung im Maß der bedingten Entropie. Es ist der Knick bei h_1, also bei Erreichen der Ordnung $p = 1$. Die Folge h_n bleibt theoretisch (also für unendlich lange Sequenzen) für alle $n \geq p = 1$ und festes ρ konstant. Der Parameter ρ erlaubt lediglich, die Unbestimmtheit des Prozesses zu variieren. Die Höhe des Plateaus (also den Wert von h_n bei großem n) bezeichnet man daher auch als die Entropie h des Prozesses (Ebeling et al., 1998). Die unterschiedlichen Plateaus der Folge h_n für $n \geq p = 1$ sind also nur ein Ausdruck der Stochastizität des Prozesses und stehen nicht direkt mit der Markov-Eigenschaft der Sequenz in Verbindung. Dabei bedeutet $\rho = 0$ maximale Unbestimmtheit (also eine Bernoulli-Sequenz) und $\rho = 1$ eine vollständig determinierte Sequenz. Das beobachtete Abfallen der Folgen h_n dieser bedingten Entropien bei größerem n ist eine Konsequenz der endlichen Sequenzlänge. Bei immer größeren Wortlängen reicht die Sequenzlänge nicht mehr aus, um alle möglichen Worte angemessen zu repräsentieren. Die Sequenz ist kein adäquates Abbild des Prozesses mehr. Es fällt auf, dass dieser Effekt bei niedrigem ρ stärker wirksam ist. Der Grund ist, dass mit Variation von ρ die Anzahl der tatsächlich auftretenden verschiedenen n-Wörter variiert.

Als Nächstes werden die Befunde aus dem Verhalten der bedingten Entropie h_n bei Markov-Sequenzen erster Ordnung auf Sequenzen höherer Ordnung übertragen. Wie oben bereits festgehalten, ist das Charakteristische einer Markov-Sequenz der Ordnung $p = 1$ der Knick im Verlauf der Folge der bedingten Entropie h_n bei h_1. Ein ähnliches Verhalten findet man auch bei Markov-Sequenzen höherer Ordnung. Die bedingte Entropie h_n ist für $n \leq p$ monoton fallend und bleibt

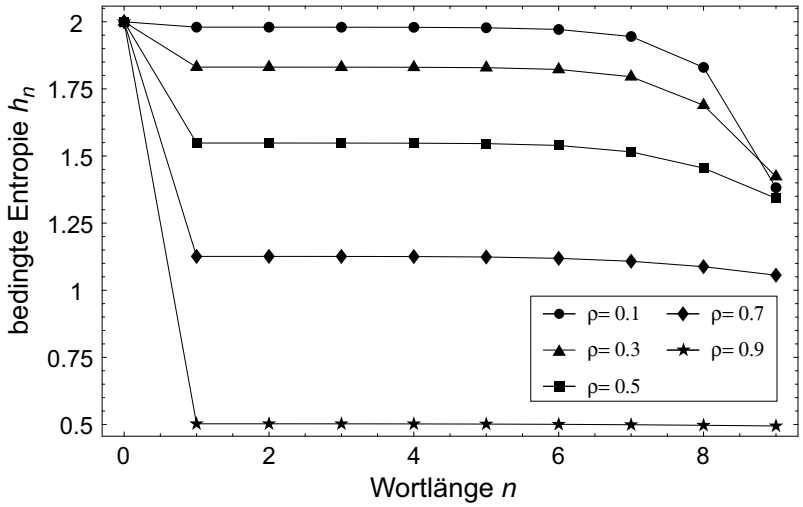

Abb. A.1. Bedingte Entropie h_n in Abhängigkeit der Wortlänge n für verschiedene Werte des Parameters ρ des DAR(1)-Prozesses aus Gleichung (A.7) mit einer Gleichverteilung als Marginalverteilung π. (Aus: Dehnert et al. (2003).)

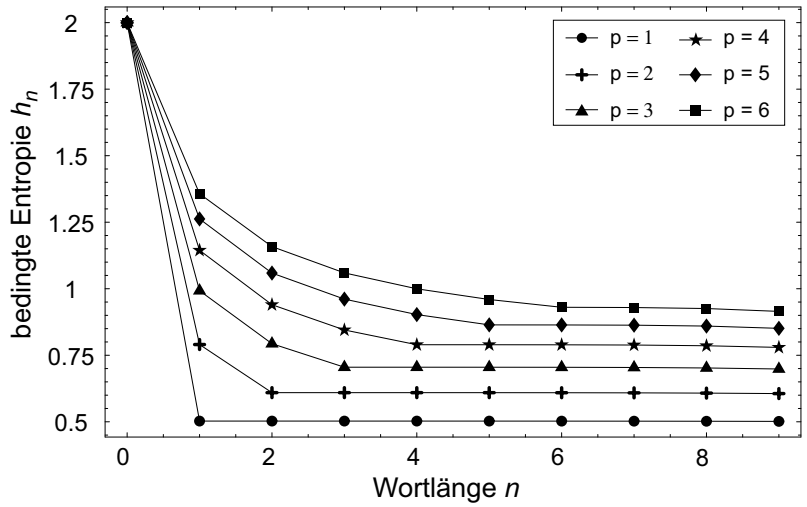

Abb. A.2. Bedingte Entropie h_n in Abhängigkeit der Wortlänge n für Markov-Sequenzen des DAR(p)-Prozesses mit $\rho = 0.9$ für $p = 1$ bis $p = 6$. Für die jeweiligen Parametervektoren $\vec{\alpha}$ und die Marginalverteilung π wurde eine Gleichverteilung gewählt. (Aus: Dehnert et al. (2003).)

nach Erreichen der Ordnung konstant, bis sie sich (bedingt durch die endliche Sequenzlänge L) bei größeren Wortlängen noch verringert. In Abbildung A.2 ist die bedingte Entropie h_n als Funktion von n für verschiedene Markov-Ordnungen p für festes $\rho = 0.9$ zu sehen. Dabei wird die vertikale Aufspaltung dieser Kurven (über die Entropie des Prozesses) vor allem durch den Parameter p bestimmt.

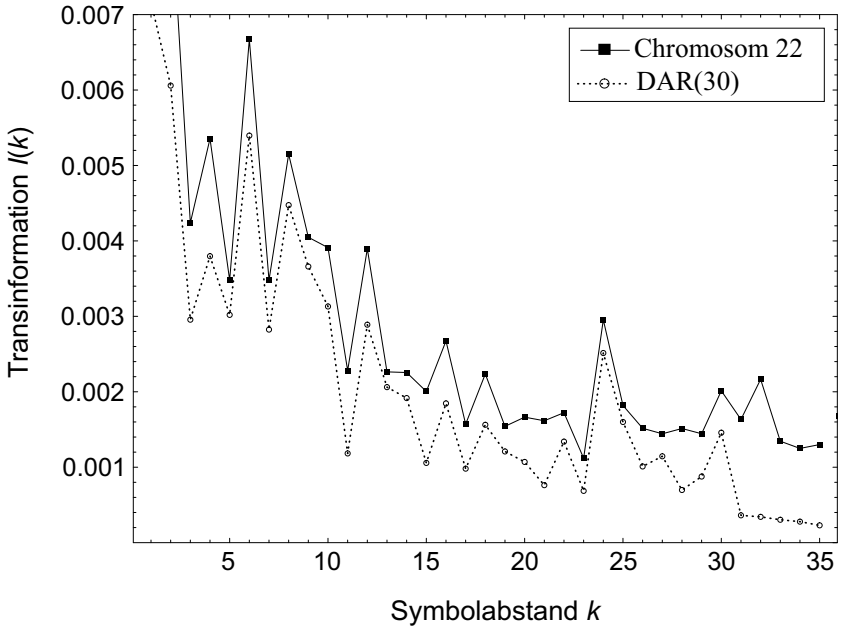

Abb. A.3. Transinformation $I(k)$ in Abhängigkeit des Symbolabstandes k für einen Parameter-Vektor $\vec{\alpha}$ des DAR(p)-Prozesses, geschätzt aus einer DNA Sequenz mit $p = 30$ (*gestrichtelte Kurve*), und tatsächlicher Verlauf der Transinformation für die reale DNA-Sequenz (*durchgezogene Kurve*). (Aus: Dehnert et al. (2003).)

Eine Parametergruppe des DAR(p)-Prozesses haben wir bisher noch nicht besprochen. Es ist der Parameter-Vektor $\vec{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_p)$. Die Komponenten α_i dieses Vektors sind, wie wir gesehen haben, (bedingte) Wahrscheinlichkeiten für ein Zurückgreifen um genau i Positionen bei der Ermittlung des nächsten Symbols in der Sequenz. Sie stellen auf diese Weise das *Gedächtnis* des Prozesses dar und bestimmen die Stärke der Korrelation zwischen den Symbolen in Abhängigkeit des Abstandes 1 bis p. Die bisher diskutierten Analysen solcher DAR(p)-Prozesse basieren auf Symbolsequenzen, deren Symbole für den Abstand 1 bis p gleich stark korreliert sind; d.h. der Vektor $\vec{\alpha}$ ist so gewählt, dass die Werte $1, \ldots, p$ mit derselben Wahrscheinlichkeit $1/p$ angenommen werden. Durch entsprechende Wahl des Parameter-Vektors $\vec{\alpha}$ lassen sich Markov-Sequenzen konstruieren, deren Korrelation mit wachsendem Abstand der Symbole abnimmt oder auch zunimmt.

Wie wir gesehen haben, ist es möglich, mit Hilfe des DAR(p)-Prozesses Symbolsequenzen einer vorgegebenen Markov-Ordnung p zu generieren, wobei die Korrelationsstärke im Abstand $i \leq p$ durch den Parametervektor $\vec{\alpha}$ festgelegt ist und die Menge an Zufall in der Sequenz über den Parameter ρ variiert werden kann. Mit Hilfe der bedingten Entropien und der Transinformation lassen sich Eigenschaften und − in gewissem Rahmen − auch Parameter eines solchen erzeugenden Prozesses aus den beobachteten Sequenzen extrahieren. Dieses Vorgehen, Sequenzen mit bekannten Prozessen zu erzeugen, um zu überprüfen, mit welchen Analyseverfahren man Zugriff auf die Prozesseigenschaften erhält, ist eine wichtige Strategie bioinformatischer Datenanalyse. Erst eine solche Validierung (*Eichung*) der Analysemethoden ermöglicht eine verlässliche Anwendung auf reale Sequenzdaten.

Der DAR(p)-Prozess kann aber auch als *Modell* für kurzreichweitige Korrelationen in DNA-Sequenzen eingesetzt werden. Zur Anpassung eines solchen Modells ist es nötig, die Korrelationsstärke zwischen zwei Nukleotiden im Abstand k in einer DNA-Sequenz zu bestimmen. Diese Parameter des DAR(p)-Prozesses werden mit Hilfe der Yule-Walker-Gleichungen (siehe Kapitel 1.2.2) geschätzt. Besitzt nun − zumindest im Abstandsbereich $k \leq p$, in dem die Korrelationen explizit durch den DAR(p)-Prozess beschrieben werden − eine mit den geschätzten Parametern simulierte Sequenz ähnliche informationstheoretische Eigenschaften wie die reale Sequenz, die zur Parameterschätzung herangezogen wurde? Wir gehen dieser Frage auf folgende Weise nach: Für eine gegebene DNA-Sequenz wird die Stärke der Korrelation zweier Nukleotide im Abstand $k \leq p$ geschätzt.[1] Neben diesen p Werten, die den Parametervektor $\vec{\alpha}$ ergeben, wird der Parameter ρ und die Verteilung der Einzelwahrscheinlichkeiten für das zufällige Ziehen eines Symbols (also die Marginalverteilung) bestimmt. Die aus der realen DNA-Sequenz geschätzten Parameter werden nun in den DAR(p)-Prozess eingesetzt, und es wird eine Symbolsequenz generiert. Diese generierte Symbolsequenz kann nun wiederum mit Hilfe der Transinformation untersucht werden. Trägt man außerdem die Transinformation für die reale DNA-Sequenz auf, so kann man die Eigenschaften beider Sequenzen anhand der Transinformation vergleichen. Abbildung A.3 zeigt die Transinformation für das menschliche Chromosom 22 für $k = 1,\ldots,35$ sowie die Transinformation für eine Realisierung eines DAR(30)-Prozesses, dessen Parameter aus dem Chromosom 22 geschätzt wurden. Wie man klar erkennt, ähneln sich die Verläufe der Transinformation bis zum Erreichen der Ordnung $p = 30$. Danach fällt die Transinformation für die mit dem DAR(p)-Prozess generierte Sequenz deutlich ab.

A.2.1 Analytische Betrachtungen

Einige Aspekte eines DAR(2)-Prozesses, für den die Paarwahrscheinlichkeiten $P(X_{n+1} = c_{n+1}, X_n = c_n)$ analytisch bestimmt werden können, sollen hier diskutiert werden. Sei $\{X_n\}$ ein stationärer DAR(p)-Prozess wie in Gleichung (A.6) definiert. Dieser Prozess wird spezifiziert durch die Marginalverteilung π, wobei für π keine Beschränkungen gelten. Unabhängig von der Marginal-Verteilung wird die Korrelationsstruktur durch $\vec{\alpha}$, ρ und Gleichung (A.6) determiniert. Dies führt zu $P(X_n = a_i) = \pi(a_i)$, $a_i \in A$ und ermöglicht die Herleitung der bedingten Wahrscheinlichkeiten mit genau p Schritten in der Bedingung (Gleichung (1.5) in Jacobs und Lewis (1983)):

$$\begin{aligned} P(X_{n+1} = c_{n+1} | X_{n-p+1} = c_{n-p+1}, \ldots, X_n = c_n) \\ = (1-\rho)\pi(c_{n+1}) + \sum_{k=1}^{p} \rho\alpha_k \delta_{c_{n+1}}(c_k) \end{aligned} \tag{A.8}$$

wobei $(c_1,\ldots,c_{n+1}) \in A^{n+1}$, $\delta_y(x) = 1$ für $x = y$ und $\delta_y(x) = 0$ für $x \neq y$ ist.

Für die Ein-Schritt Übergangswahrscheinlichkeiten gilt dann:

$$\begin{aligned} P(X_{n+1} = c_{n+1} | X_n = c_n) = (1-\rho)\pi(c_{n+1}) + \rho\alpha_1 \delta_{c_{n+1}}(c_n) + \\ + \sum_{k=2}^{p} \rho\alpha_k \frac{P(X_n = c_n, X_{n-k+1} = c_{n+1})}{P(X_n = c_n)} \,. \end{aligned} \tag{A.9}$$

[1] Anschaulich kann man sich eine solche Korrelationsstärke als (systematische) Abweichung von einer Gleichverteilung des zweiten Symbols bei Vorliegen des ersten Symbols vorstellen.

Für $p = 2$ erhält man ein lineares Gleichungssystem für die Paarwahrscheinlichkeiten $P(X_{n+1} = c_{n+1}, X_n = c_n)$ durch Multiplikation von Gleichung (A.9) mit $P(X_n = c_n) = \pi(c_n)$

$$\begin{aligned} P(X_{n+1} = c_{n+1}, X_n = c_n) &= (1-\rho)\pi(c_{n+1})\pi(c_n) + \\ &+ \rho\alpha_1\delta_{c_{n+1}}(c_n)\pi(c_n) + \\ &+ \rho\alpha_2 P(X_n = c_{n+1}, X_{n+1} = c_n) \end{aligned} \tag{A.10}$$

mit $c_n, c_{n+1} \in A$.

Die Lösung dieses Gleichungssystems für $P(X_{n+1} = c_{n+1}, X_n = c_n)$ ist dann gegeben durch

$$P(X_{n+1} = c_{n+1}, X_n = c_n) = (1-\rho)\pi(c_{n+1})\pi(c_n)\frac{1}{1-\rho\alpha_2}\,, c_{n+1} \neq c_n \tag{A.11}$$

$$\begin{aligned} P(X_{n+1} = c_{n+1}, X_n = c_n) &= (1-\rho)\pi(c_{n+1})\pi(c_n)\frac{1}{1-\rho\alpha_2} + \\ &+ \pi(c_n)\frac{\rho\alpha_1}{1-\rho\alpha_2}\,, c_{n+1} = c_n\,. \end{aligned} \tag{A.12}$$

Ein Vergleich der analytisch gewonnenen Ausdrücke mit den Ergebnissen einer Simulation zeigt eine sehr gute Übereinstimmung und bestätigt somit die Resultate.

B

Ergänzende Abbildungen

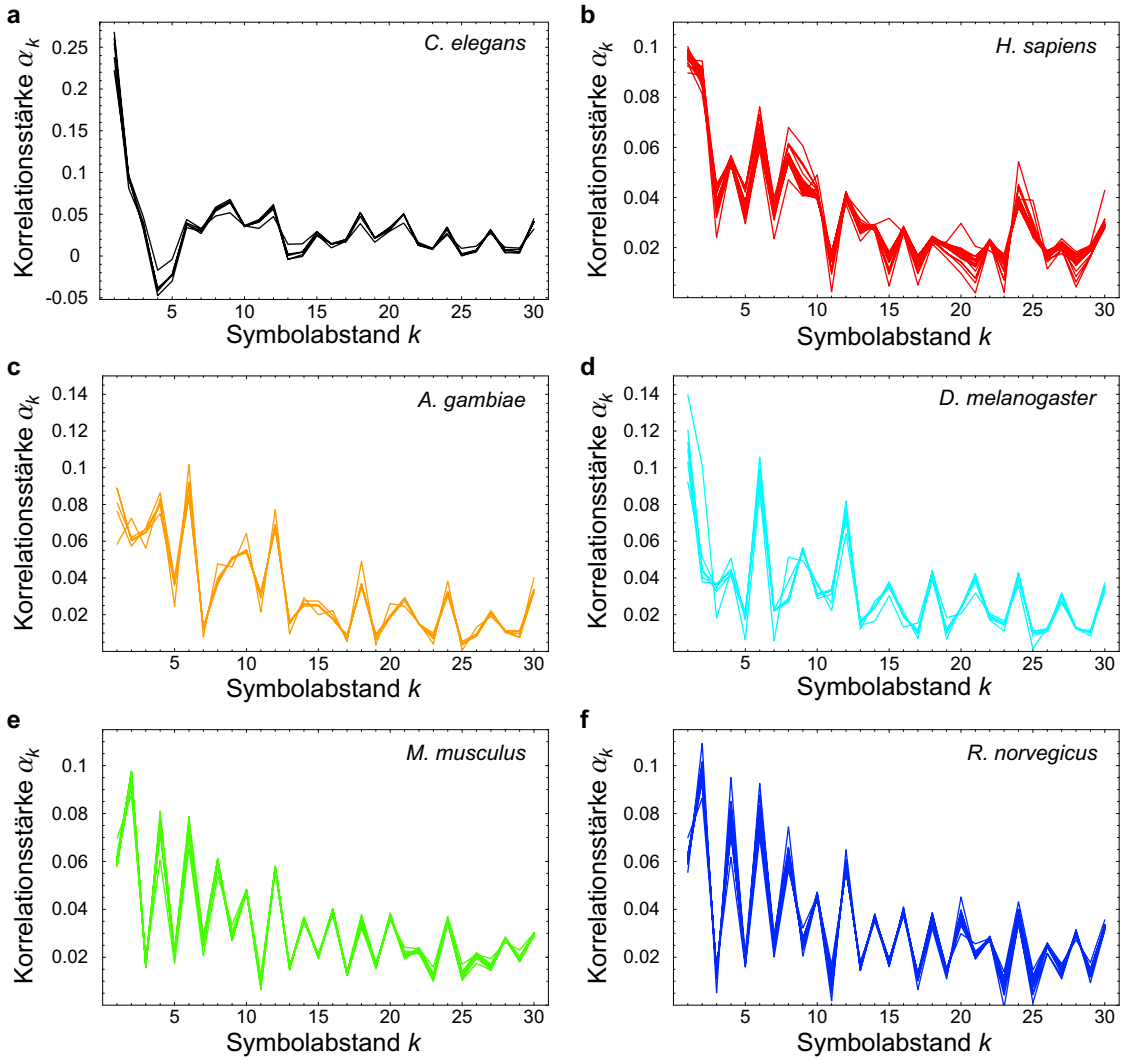

Abb. B.1. Korrelationskurven der Markov-Repräsentation für $p = 30$ für die Chromosomen der folgenden Spezies: **a** *C. elegans* [6 Kurven], **b** *H. sapiens* [24 Kurven], **c** *A. gambiae* [5 Kurven], **d** *D. melanogaster* [6 Kurven], **e** *M. musculus* [20 Kurven] und *R. norvegicus* [21 Kurven].

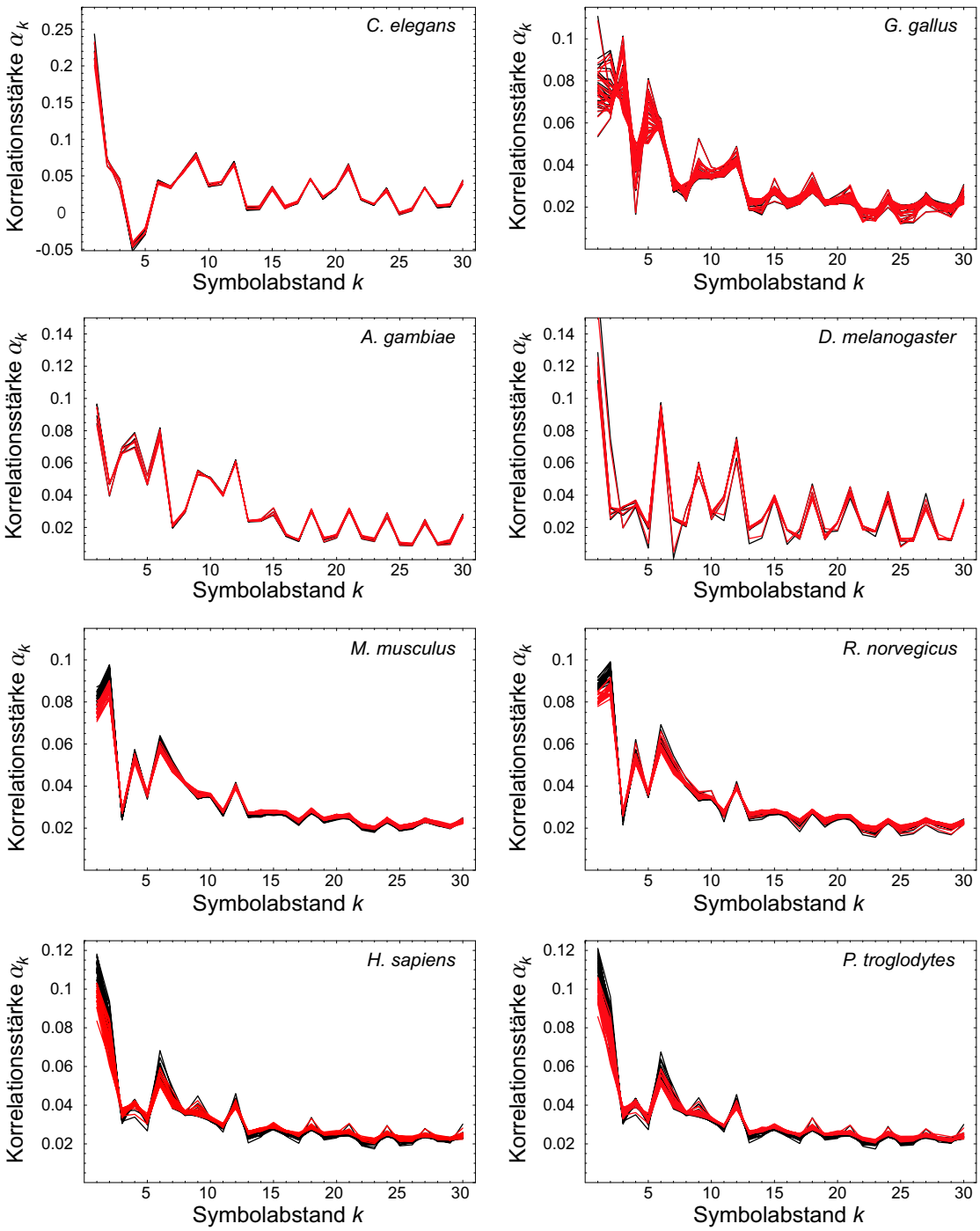

Abb. B.2. Korrelationskurven nach der Maskierung aller repetitiver Elemente für zwei unterschiedliche Maskierungsverfahren. In schwarz dargestellt sind die Korrelationskurven der einzelnen Spezies, die man durch Ausschneiden der repetitiven Elemente erhält. Das Überschreiben der repetitiven Elemente mit zufälligen Symbolssequenzen führt zu den für jede Spezies in rot dargestellten Korrelationskurven.

C

Datenquellen

Tabelle C.1. Quellen der intern angelegten Genom-Datenbank DNA_DATA_01.

Spezies	**Datenquelle und** **Veröffentlichung/*Accession-Number***
1. *Anopheles gambiae*	ftp://ftp.ensembl.org anopheles-21.2b
2. *Arabidopsis thaliana*	ftp://ftp.ncbi.nih.gov/genomes/ Arabidopsis_thaliana/ (Date: 23.05.03)
3. *Caenorhabditis elegans*	ftp://ftp.ensembl.org celegans-21.116a
4. *Cryptosporidium parvum*	http://www.ncbi.nlm.nih.gov BX526834
5. *Drosophila melanogaster*	ftp://ftp.ensembl.org fly-21.3a
6. *Encephalitozoon cuniculi*	http://www.ebi.ac.uk/genomes AL391737, AL590442 - AL590451 AE016814 - AE016820
7. *Homo sapiens*	ftp://ftp.ensembl.org human-21.34d
8. *Leishmania major*	http://www.ebi.ac.uk/genomes AE001274, AC125735
9. *Mus musculus*	ftp://ftp.ensembl.org mouse-21.32b
10. *Oryza sativa*	http://www.ebi.ac.uk/genomes BA000010, BA000044
11. *Plasmodium falciparum*	http://www.ncbi.nlm.nih.gov NC_004325, NC_000910, NC_000521, NC_004318, NC_004326, NC_004327, NC_004328, NC_004329, NC_004330, NC_004314, NC_004315, NC_004316, NC_004331, NC_004317
12. *Rattus norvegicus*	ftp://ftp.ensembl.org rat-21.3b
13. *Saccharomyces cerevisiae*	http://www.ncbi.nlm.nih.gov NC_001133 - NC_001148
14. *Schizosaccharomyces pombe*	http://www.ncbi.nlm.nih.gov NC_003424, NC_003423, NC_003421
15. *Trypanosoma brucei*	http://www.ebi.ac.uk/genomes AL929608

Tabelle C.2. Quellen der intern angelegten Genom-Datenbank DNA_DATA_02.

Spezies	**Datenquelle und** **Veröffentlichung/*Accession-Number***
1. *Anopheles gambiae*	ftp://ftp.ensembl.org anopheles-22.2b
2. *Arabidopsis thaliana*	http://www.ncbi.nlm.nih.gov NC_003070, NC_003071, NC_003074, NC_003075, NC_003076
3. *Ashbya gossypii*	http://www.ebi.ac.uk/genomes
4. *Caenorhabditis elegans*	ftp://ftp.ensembl.org celegans-22.116a
5. *Danio rerio*	ftp://ftp.ensembl.org zebrafish-22.3b
6. *Drosophila melanogaster*	ftp://ftp.ensembl.org fly-22.3a
7. *Encephalitozoon cuniculi*	http://www.ebi.ac.uk/genomes AL391737, AL590442 - AL590451 AE016814 - AE016820
8. *Gallus gallus*	ftp://ftp.ensembl.org chicken-22.1
9. *Homo sapiens*	ftp://ftp.ensembl.org human-22.34d
10. *Mus musculus*	ftp://ftp.ensembl.org mouse-22.32b
11. *Pan troglodytes*	ftp://ftp.ensembl.org chimp-22.1
12. *Plasmodium falciparum*	http://www.ncbi.nlm.nih.gov NC_004325, NC_000910, NC_000521, NC_004318, NC_004326, NC_004327, NC_004328, NC_004329, NC_004330, NC_004314, NC_004315, NC_004316, NC_004331, NC_004317
18. *Rattus norvegicus*	ftp://ftp.ensembl.org rat-22.3b
19. *Saccharomyces cerevisiae*	http://www.ncbi.nlm.nih.gov NC_001133 - NC_001148

Tabelle C.3. Quellen der intern angelegten Genom-Datenbank DNA_DATA_03.

Spezies	**Datenquelle und** **Veröffentlichung/*Accession-Number***
1. *Anopheles gambiae*	http://genome.ucsc.edu/ anoGam1 (IAGP v.MOZ2)
2. *Caenorhabditis elegans*	http://genome.ucsc.edu/ ce2 (WormBase v. WS120)
3. *Drosophila melanogaster*	http://genome.ucsc.edu/ dm2 (BDGP Release 4)
4. *Gallus gallus*	ftp://ftp.ensembl.org galGal2 (WUSTL Feb. 2004 release)
5. *Homo sapiens*	http://genome.ucsc.edu/ hg17 (NCBI Build 35)
6. *Mus musculus*	http://genome.ucsc.edu/ mm8 (NCBI Build 36)
7. *Pan troglodytes*	http://genome.ucsc.edu/ panTro1 (CGSC Build 1 Version 1)
8. *Rattus norvegicus*	http://genome.ucsc.edu/ rn3 (Baylor College of Medicine HGSC v3.1)

Tabelle C.4. Auflistung der den Abbildungen in dieser Arbeit zugrunde liegenden Datensätze. Die Angabe zum Datensatz bezieht sich dabei auf die Tabellen C.1, C.2 und C.3.

Abbildung	Datensatz	Abbildung	Datensatz
2.1	DNA_DATA_01	2.15	DNA_DATA_02
2.2	DNA_DATA_01	2.16	DNA_DATA_02
2.3	DNA_DATA_01	2.17	DNA_DATA_02
2.4	DNA_DATA_01	2.18	DNA_DATA_02
2.5	DNA_DATA_01	2.19	DNA_DATA_02
2.6	DNA_DATA_01	2.20	DNA_DATA_02
2.7	DNA_DATA_01	2.21	DNA_DATA_02
2.8	DNA_DATA_01	2.22	DNA_DATA_02
2.9	DNA_DATA_01	2.24	DNA_DATA_03
2.10	DNA_DATA_02	2.25	DNA_DATA_03
2.11	DNA_DATA_02	2.27	DNA_DATA_03
2.12	DNA_DATA_02	2.28	DNA_DATA_03
2.13	DNA_DATA_02	2.29	DNA_DATA_03
2.14	DNA_DATA_02	B.1	DNA_DATA_02
		B.2	DNA_DATA_02

Tabelle C.5. Alte und neue Bezeichnungsweise der Chromosomen des Schimpansen (*Pan troglodytes*) im Vergleich zur Benennung der menschlichen Chromosomen. (Aus http://www.ensembl.org/Pan_troglodytes/chromosomes.html)

Menschliche Chromosomenbezeichnung	**Neue Einteilung des Schimpansen**	**Alte Einteilung des Schimpansen**
1	1	1
2p-q13	2A	12
2q-qter	2B	13
3	3	2
4	4	3
5	5	4
6	6	5
7	7	6
8	8	7
9	9	11
10	10	8
11	11	9
12	12	10
13	13	14
14	14	15
15	15	16
16	16	18
17	17	19
18	18	17
19	19	20
20	20	21
21	21	22
22	22	23
X	X	X
Y	Y	Y

Literaturverzeichnis

Almeida, P., Penha-Goncalves, C., 2004. Long perfect dinucleotide repeats are typical of vertebrates, show motif preferences and size convergence. Mol. Biol. Evol. 21, 1226–1233.

Arndt, P., Burge, C., Hwa, T., 2002. DNA sequence evolution with neighbor-dependent mutation. In: Proceedings of the 6th Annual International Conference on Computational Biology, (RECOMB 2002). Washington DC. ACM Press, New York, pp. 32–38.

Arndt, P. F., Hwa, T., 2005. Identification and measurement of neighbor-dependent nucleotide substitution processes. Bioinformatics 21, 2322–2328.

Batzer, M. A., Deininger, P. L., 2002. Alu repeats and human genomic diversity. Nature Reviews Genetics 3, 370–379.

Beckman, J. S.and Weber, J. L., 1992. Survey of human and rat microsatellites. Genomics 12, 627–631.

Benson, G., 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucl. Acids Res. 27, 573–580.

Bernardi, G., 1989. The Isochore Organization of the Human Genome. Annual Review of Genetics 23, 637–659.

Bernardi, G., 2000. Isochores and the evolutionary genomics of vertebrates. Gene 241, 3–17.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., Rodier, F., 1985. The mosaic genome of warm-blooded vertebrates. Science 228, 953–958.

Castelo, A. T., Martins, W., Gao, G. R., 2002. TROLL–Tandem Repeat Occurrence Locator. Bioinformatics 18, 634–636.

Celniker, S., Wheeler, D., Kronmiller, B., et al., 2002. Finishing a whole-genome shotgun: Release 3 of the Drosophila melanogaster euchromatic genome sequence. Genome Biology 3, research0079.1–0079.14.

Chatzidimitriou-Dreismann, C. A., Larhammar, D., 1993. Long-range correlations in DNA. Nature 361, 212–213.

Dehnert, M., Helm, W. E., Hütt, M.-T., 2003. A discrete autoregressive process as a model for short-range correlations in DNA sequences. Physica A 327, 535–553.

Dehnert, M., Helm, W. E., Hütt, M.-T., 2005a. Information theory reveals large-scale synchronisation of statistical correlations in Eukaryote genomes. Gene 345, 81–90.

Dehnert, M., Helm, W. E., Hütt, M.-T., 2006. The informational structure of two closely related eukaryotic genomes. Phys. Rev. E, eingereicht.

Dehnert, M., Plaumann, R., Helm, W. E., Hütt, M.-T., 2005b. Genome phylogeny based on short-range correlations in DNA sequences. J. Comp. Biol. 12, 545–553.

Deininger, P. L., Batzer, M. A., 2002. Mammalian Retroelements. Genome Res. 12 (10), 1455–1465.

Deininger, P. L., Morany, J. V., Batzer, M. A., Kazazian Jr, H. H., 2003. Mobile elements and mammalian genome evolution. Current Opinion in Genetics & Development 13, 651–658.

Dewannieux, M., Esnault, C., Heidmann, T., 2003. LINE-mediated retrotransposition of marked Alu sequences. Nat. Genet. 35, 41–48.

Ebeling, W., Freund, J., Schweitzer, F., 1998. Komplexe Strukturen: Entropie und Information. Teubner, Stuttgart.

Efron, B., Tibshirani, R., 1993. An Introduction to the Bootstrap. Chapman&Hall/CRC, Boca Raton/FL.

Ellegren, H., 2004. Microsatellites: simple sequences with complex evolution. Nature Reviews Genetics 5, 435–445.

Gentles, A. J., Karlin, S., 2001. Genome-scale compositional comparisons in eukaryotes. Genome Res. 11, 540–546.

Grosse, I., Herzel, H., Buldyrev, S. V., Stanley, H. E., 2000. Species independence of mutual information in coding and noncoding DNA. Phys. Rev. E 61, 5624–5629.

Hameister, J., 2006. Zur effizienten Implementierung von DAR(p)-Prozessen. Unveröffentlicht.

Hao, B., Qi, J., 2003. Prokaryote Phylogeny without Sequence Alignment: From Avoidance Signature to Composition Distance. IEEE Proceedings of the Computational Systems Bioinformatics.

Hedges, D. J., Batzer, M. A., 2005. From the margins of the genome: mobile elements shape primate evolution. BioEssays 27, 785–794.

Hedges, S. B., 2002. The origin and evolution of model organisms. Nat. Rev. Genet. 3, 838–849.

Herzel, H., Ebeling, W., 1985. The decay of correlations in chaotic maps. Phys. Lett. A 111, 1–4.

Herzel, H., Grosse, I., 1995. Measuring correlations in symbolic sequences. Physica A 216, 518–542.

Herzel, H., Grosse, I., 1997. Correlations in DNA sequences: The role of protein coding segments. Phys. Rev. E 55, 800–809.

Herzel, H., Weiss, O., Trifonov, E. N., 1999. 10-11 bp periodicities in complete genomes reflect protein structure and DNA folding. Bioinformatics 15, 187–193.

Hinrichs, A. S., Karolchik, D., Baertsch, R., et al., 2006. The UCSC Genome Browser Database: update 2006. Nucl. Acids Res. 34, D590–598.

Holste, D., Grosse, I., Beirer, S., Schieg, P., Herzel, H., 2003. Repeats and correlations in human DNA sequences. Phys. Rev. E 061913, 1–9.

Hütt, M.-T., Dehnert, M., 2006. Methoden der Bioinformatik. Eine Einführung. Springer, Berlin Heidelberg New York.

Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Human Genome Sequencing Consortium, 2004. Finishing the euchromatic sequence of the human genome. Nature 431, 931–945.

International Chicken Genome Sequencing Consortium, 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432, 695–716.

Jacobs, P., Lewis, P., 1978. Discrete time series generated by mixtures III: autoregressive processes (DAR(p)). Tech. Rep. NPS55-78-022, Naval Postgraduate School, Monterey, California.

Jacobs, P., Lewis, P., 1983. Stationary discrete autoregressive-moving average time series generated by mixtures. Journal of Time Series Analysis 4, 19–36.

Jurka, J., Kapitonov, V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J., 2005. Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110, 462–467.

Jurka, J., Klonowski, P., Dagman, V., Pelton, P., 1996. CENSOR - a program for identification and elimination of repetitive elements from DNA sequences. Computers and Chemistry 20, 119–122.

Karlin, S., 1998. Global dinucleotide signatures and analysis of genomic heterogeneity. Curr. Opin. Microbiol. 1, 598–610.

Karlin, S., Brendel, V., 1993. Patchiness and correlations in DNA sequences. Science 259, 677–680.

Karlin, S., Ladunga, I., 1994. Comparisons of Eukaryotic Genomic Sequences. PNAS 91 (26), 12832–12836.

Karlin, S., Mrázek, J., 1997. Compositional differences within and between eukaryotic genomes. PNAS 94, 10227–10232.

Karlin, S., Taylor, H. M., 1975. A first course in stochastic processes, 2nd Edition. Academic Press Inc.(New York) Ltd.

Kaufman, L., Rousseeuw, P., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, New York Chichester.

Kazazian, H. J., 2004. Mobile elements: drivers of genome evolution. Science 303, 1626–1632.

Krauss, R., 2006. Klassifikation von Spezies durch Korrelationssignaturen in genomweiten DNA-Sequenzen: Vergleich von Längenskalen mit Hilfe des SAS-Systemes. Diplomarbeit, FH Darmstadt.

Kunkel, T. A., Bebenek, K., 2000. DNA Replication Fidelity. Annual Review of Biochemistry 69, 497–529.

Li, W., 1989. Spatial 1/f spectra in open dynamical systems. Europhys. Lett. 10, 395–400.

Li, W., 1991. Expansion-modification systems: A model for spatial $1/f$ spectra. Phys. Rev. A 43, 5240–5260.

Li, W., Holste, D., 2004a. An unusual 500,000 bases long oscillation of guanine and cytosine content in human chromosome 21. Computational Biology and Chemistry 28, 393–399.

Li, W., Holste, D., 2004b. Spectral analysis of guanine and cytosine fluctuation of mouse genomic DNA. Fluctuation and Noise Letters 4, L453–L464.

Li, W., Holste, D., 2005. Universal 1/f noise, crossovers of scaling exponents, and chromosome-specific patterns of guanine-cytosine content in DNA sequences of the human genome. Phys. Rev. E 71, 041910.

Li, W., Kaneko, K., 1992. Long-range correlation and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence. Europhys. Lett. 17, 655–660.

Li, W., Marr, T. G., Kaneko, K., 1994. Understanding long-range correlations in DNA sequences. Physica D 82, 392–416.

Li, Y.-C., Korol, A. B., Fahima, T., Nevo, E., 2004. Microsatellites Within Genes: Structure, Function, and Evolution. Mol. Biol. Evol. 21, 991–1007.

Luning Prak, E. T., Kazazian, H. H., 2000. Mobile elements and the human genome. Nat. Rev. Genet. 1, 134–144.

Maddox, J., 1992. Long-range correlations within DNA. Nature 358, 103–103.

McConkey, E., 2004. Orthologous numbering of great ape and human chromosomes is essential for comparative genomics. Cytogenetic and Genome Research 105, 157–158.

Messer, P. W., Arndt, P. F., Lässig, M., 2005. Solvable Sequence Evolution Models and Genomic Correlations. Phys. Rev. Lett. 94, 138103.

Mouse Genome Sequencing Consortium, 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562.

Murnane, J. P., Morales, J. F., 1995. Use of a mammalian interspersed repetitive (MIR) element in the coding and processing sequences of mammalian genes. Nucl. Acids Res. 23, 2837–2839.

Nee, S., 1992. Uncorrelated DNA walks. Nature 357, 450–450.

Nei, M., Kumar, S., 2000. Molecular Evolution and Phylogenetics. Oxford Univ Press, New York.

Peng, C., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E., Goldberger, A. L., 1994. Mosaic organization of DNA nucleotides. Phys. Rev. E 49, 1685–1689.

Peng, C.-K., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Sciortino, F., Simons, M., Stanley, H. E., 1992. Long-range correlations in nucleotide sequences. Nature 356, 168–170.

Plaumann, R., 2003. Über die Speziesabhängigkeit von kurzreichweitigen Korrelationen in DNA-Sequenzen. Diplomarbeit, FH Darmstadt.

Qi, J., Wang, B., Hao, B., 2004. Whole genome prokaryote phylogeny without sequence alignment: a K-string composition approach. J. Mol. Evol. 58, 1–11.

Rat Genome Sequencing Project Consortium, 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428, 493–521.

Russell, G., Subak-Sharpe, J., 1977. Similarity of the general designs of protochordates and invertebrates. Nature 266, 533–536.

Russell, G., Walker, P., Elton, R., Subak-Sharpe, J., 1976. Doublet frequency analysis of fractionated vertebrate nuclear DNA. J. Mol. Biol. 108, 1–23.

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4, 406–425.

Shannon, C., 1948. A mathematical theory of communication. Bell System Technical Journal 27, 379–423,623–656.

Smit, A., Hubley, R., Green, P., 2004. RepeatMasker Open-3.0. at http://www.repeatmasker.org.

Sokal, R., Sneath, P., 1963. Principles of numerical taxonomy. W.H. Freeman and Company.

Takai, D., Jones, P., 2002. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. PNAS 99, 3740–3745.

The Arabidopsis Genome Initiative, 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408, 796–815.

The C. elegans Sequencing Consortium, 1998. Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. Science 282, 2012–2018.

The Chimpanzee Sequencing and Analysis Consortium, 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437, 69–87.

Trifonov, E. N., Sussman, J. L., 1980. The pitch of chromatin DNA is reflected in its nucleotide sequence. PNAS 77, 3816–3820.

Venter, C. J., Adams, M. D., Myers, E. W., et al., 2001. The Sequence of the Human Genome. Science 291, 1304–1351.

Voss, R. F., 1992. Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences. Phys. Rev. Lett. 68, 3805–3808.

Webster, M. T., Smith, N. G. C., Ellegren, H., 2002. Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments. PNAS 99, 8748–8753.

Danksagung

Professor Marc-Thorsten Hütt möchte ich für sein großes Interesse an diesem Promotionsprojekt danken. Seine immerwährende Diskussionsbereitschaft und Unterstützung waren mir eine große Hilfe.

Professor Felicitas Pfeifer danke ich für die Möglichkeit, an Vorlesungen und Praktika der Säule Mikrobiologie teilzunehmen, für ihr Interesse an dieser Arbeit und die Übernahme des 1. Korreferats.

Professor Werner E. Helm danke ich für Übernahme des 2. Korreferats und die gemeinsamen Diskussionen und Reflexionen zu dem hier behandelten Themengebiet und darüber hinaus.

Rainer Plaumann und Jörn Hameister gilt mein Dank für die Unterstützung bei der Entwicklung einer geschwindigkeitsoptimierten Umsetzung der Softwarewerkzeuge in C++.

Ich danke Stefan Christ, Heike Hameister und Rainer Plaumann für das aufmerksame Korrekturlesen dieser Arbeit.

Der Arbeitsgruppe Hütt danke ich für die schöne Zeit, die angenehmen Gespräche und den freundschaftlichen Umgang.

Lebenslauf

von Manuel Dehnert, geboren am 03.03.1977 in Bad Hersfeld, verheiratet

Schule:

1984 - 1987:	Grundschule in Bad Hersfeld
1987 - 1993:	Realschule in Bad Hersfeld
1993 - 1996:	Fachoberschule Bad Hersfeld, Schwerpunkt Informationstechnik

Studium

1997 - 2002:	Mathematik, Fachhochschule Darmstadt Schwerpunkte: Statistik, Informatik, Physik
	Diplomarbeit in Kooperation mit dem Fachbereich Biologie der Technischen Universität Darmstadt Titel der Arbeit: Untersuchungen zum Einsatz von informationstheoretischen Maßen bei der Analyse von DNA-Sequenzen
	Berufspraktische Semester bei Opel Antwerpen (Belgien) und Helaba London (England)

Praktische Tätigkeiten

1996 - 1997:	Zivildienst in der Jugendwerkstatt Bad Hersfeld e.V.
09/2002 - 02/2003:	Freier Mitarbeiter der Berlin-Brandenburgischen Akademie der Wissenschaften
10/2002 - 02/2003:	Wissenschaftliche Hilfskraft der Technischen Universität Darmstadt, Mikrobiologie und Genetik (GK), FB10

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Dissertation selbstständig und nur mit den angegebenen Hilfsmitteln angefertigt habe.

Darmstadt, den 22.05.06